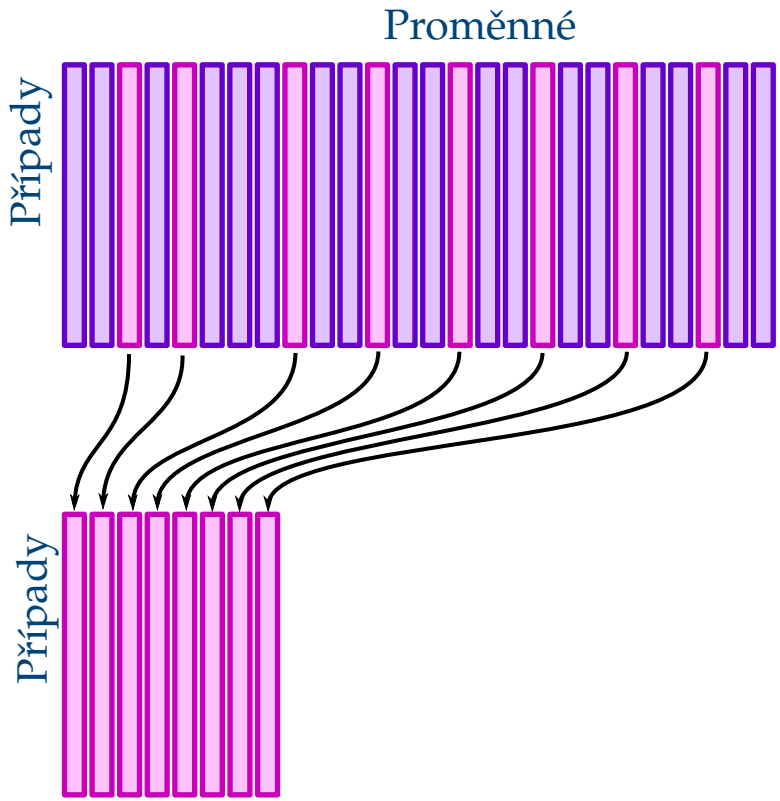
Selekce a extrakce příznaků.

Petr Pošík

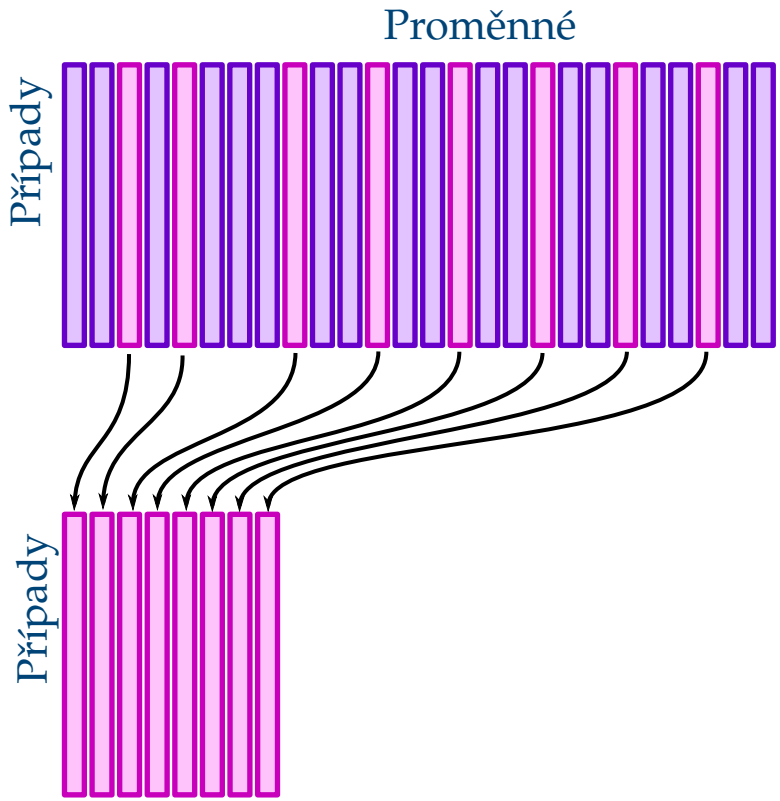
Katedra kybernetiky
ČVUT FEL

Selekce a extrakce příznaků

Proč?



Proč?



Tisíce až miliony veličin (příznaků, atributů): výběr těch nejdůležitějších pro konstrukci

- ✔ přesnějších,
- ✔ rychlejších a
- ✔ jednodušších modelů.

Příklad: data o kosatcích (Iris)

Úplný průzkum všech kombinací vstupů (Weka klasifikátory `J48` a `IBk`):
LOO Xval Error: Leave-one-out crossvalidation error

Přístup	Vstupní proměnné SL	SW	PL	PW	LOO Xval Error J48	3-NN
Žádné vstupy					100.0 %	100.0 %
1 vstup	x				26.7 %	28.7 %
		x			41.3 %	47.3 %
			x		6.0 %	8.0 %
				x	5.3 %	4.0 %
2 vstupy	x	x			23.3 %	24.0 %
	x		x		6.7 %	5.3 %
	x			x	5.3 %	4.0 %
		x	x		6.0 %	6.0 %
		x		x	5.3 %	4.7 %
			x	x	4.7 %	5.3 %
3 vstupy	x	x	x		6.7 %	7.3 %
	x	x		x	5.3 %	5.3 %
	x		x	x	4.7 %	3.3 %
		x	x	x	4.7 %	4.7 %
Všechny vstupy	x	x	x	x	4.7 %	4.7 %

- ✔ J48: dosahuje nejmenší chyby (4.7 %) vždy, když jsou mezi vstupy PL a PW; je schopná si tyto dvě proměnné vybrat sama, proměnné navíc jí nevadí.
- ✔ 3-NN: sama neobsahuje žádnou metodu selekce proměnných, pracuje se všemi, které jí dáme. Nejlepší výsledek zpravidla *nedosahuje* při použití všech vstupů!

Nomenklatura

Metody z hlediska počtu proměnných posuzovaných najednou:

- ✔ **Jednorozměrmé (univariate, variable ranking)**:
  posuzují proměnné (atributy) jednu po druhé
- ✔ **Mnoharozměrné (multivariate, variable subset selection)**:
  posuzují celé skupiny proměnných najednou

Nomenklatura

Metody z hlediska počtu proměnných posuzovaných najednou:

- ✔ **Jednorozměrmé (univariate, variable ranking)**:
  posuzují proměnné (atributy) jednu po druhé
- ✔ **Mnoharozměrné (multivariate, variable subset selection)**:
  posuzují celé skupiny proměnných najednou

Metody z hlediska využití modelu (metody učení), který bude následně na data použit:

- ✔ **Filter**: vybírá podskupinu proměnných nezávisle na modelu
- ✔ **Wrapper**: vybírá podskupinu proměnných s ohledem na model
- ✔ **Embedded method**: model má metodu výběru vstupních proměnných přímo zabudovánu v sobě (např. klasifikační a regresní stromy)

Jednorozměrné metody výběru proměnných

Jednorozměrné metody (variable ranking)

- ✔ hlavní nebo pomocná technika v mnoha složitějších metodách
- ✔ jednoduchá a škálovatelná metoda, v praxi často dobře funguje
- ✔ kritéria řazení definovaná pro jednotlivé vstupní veličiny, nezávisle na kontextu daném ostatními vstupními veličinami

Možné kombinace typů vst. a výst. veličiny (nekompletní výčet použitelných metod)

Vstup. proměnná X	Výstupní proměnná Y: Nominální	Výstupní proměnná Y: Spojitá
Nominální	Analýza kontingenční tabulky $p(Y)$ vs. $p(Y\|X)$ χ^2-test nezávislosti Inf. zisk (viz klasifikační stromy)	T-test, ANOVA ROC (AUC) diskretizace Y (viz levý sloupec)
Spojitá	T-test, ANOVA ROC (AUC) logistická regrese diskretizace X (viz horní řádek)	korelace regrese diskretizace Y (viz levý sloupec) diskretizace X (viz horní řádek)

- ✔ všechny metody poskytují skóre, pomocí něhož lze proměnné seřadit podle "velikosti vlivu" na závislou proměnnou
- ✔ statistické testy poskytují o tzv. p-hodnoty, neboli dosaženou hladinu významnosti; ty mohou sloužit pro posouzení "důležitosti" atributu v absolutním měřítku

Dále uvidíte…

Výběr metod pro posouzení souvislosti 2 proměnných:

- ✔ Korelační kritéria
- ✔ Klasifikátory podle 1 vstupní proměnné
- ✔ Míry založené na teorii informace
- ✔ Statistické míry

Korelační kritéria

- *Rozptyl (variance)* náhodné veličiny popisuje, jaký "rozsah" veličina zabírá.

$$Var(X) = s_X^2 = \frac{1}{n-1}\sum(X - \bar{X})^2 = Cov(X, X)$$

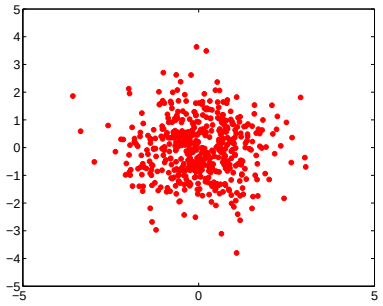
- *Kovariance* dvou náhodných veličin popisuje, jak se dvě veličiny společně mění.

$$Cov(X, Y) = s_{XY} = \frac{1}{n-1}\sum(X - \bar{X})(Y - \bar{Y})$$

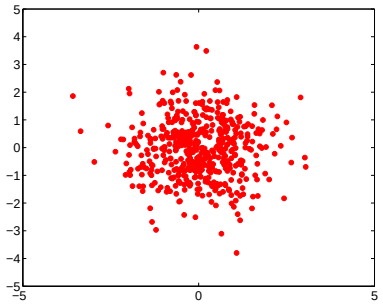
- *Korelace* měří sílu závislosti mezi dvěma spojitými veličinami. Pro měření síly *lineární* závislosti se používá *Pearsonův korelační koeficient*:

$$r_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} \in \langle -1, 1 \rangle$$

Korelace: příklady

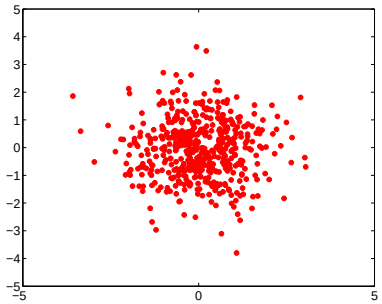


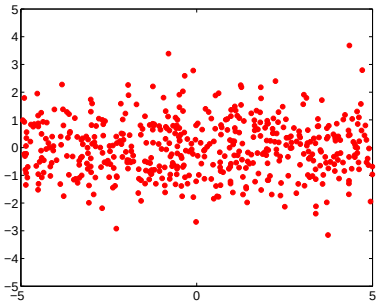
Korelace: příklady



$r = 0$

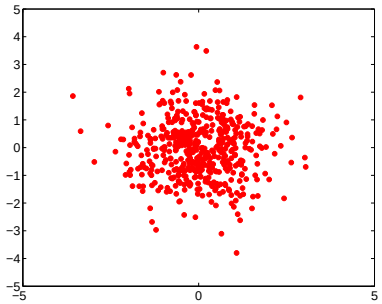
Korelace: příklady



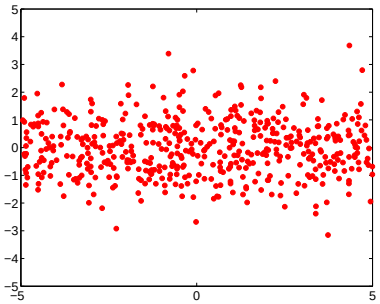


$r = 0$

Korelace: příklady

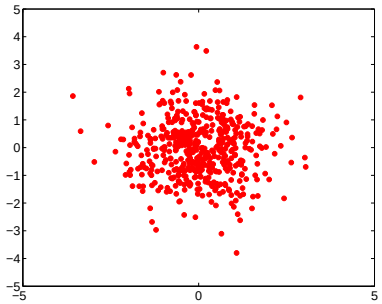


$r = 0$

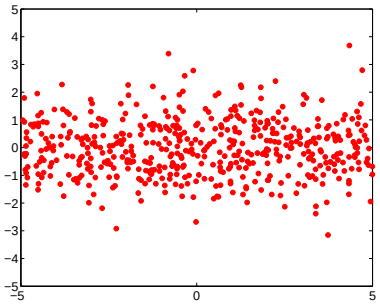


$r = 0$

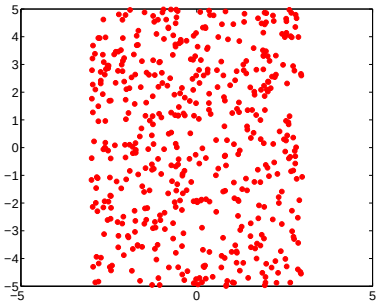
Korelace: příklady



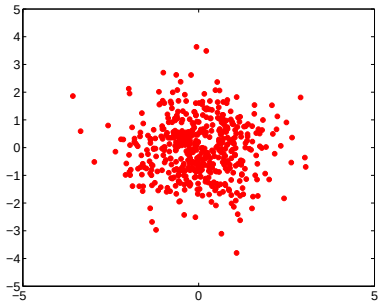
$r = 0$



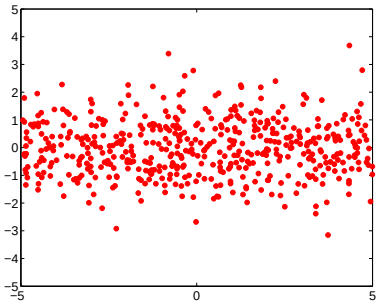
$r = 0$



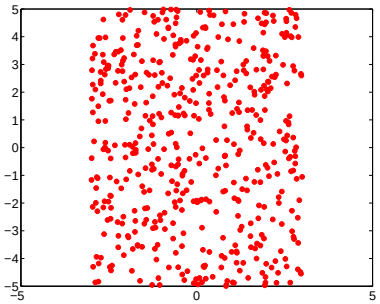
Korelace: příklady



$r = 0$

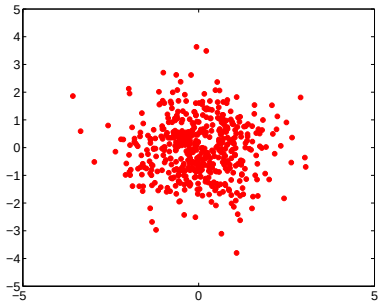


$r = 0$

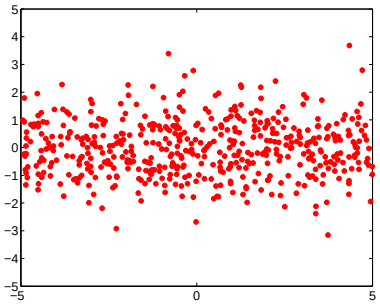


$r = 0$

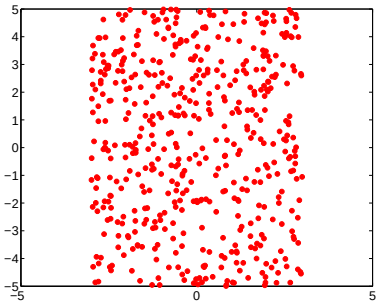
Korelace: příklady



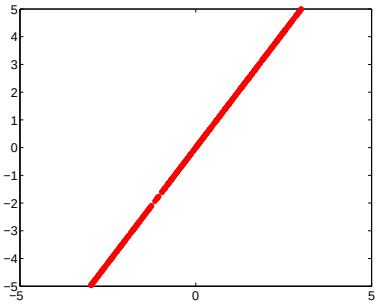
$r = 0$



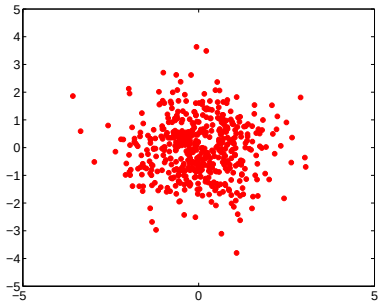
$r = 0$



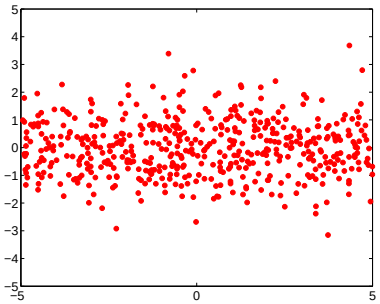
$r = 0$



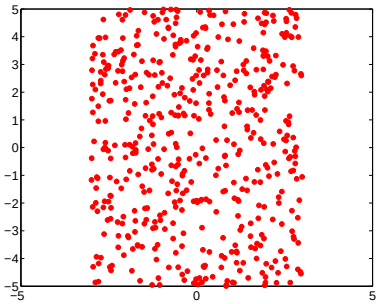
Korelace: příklady



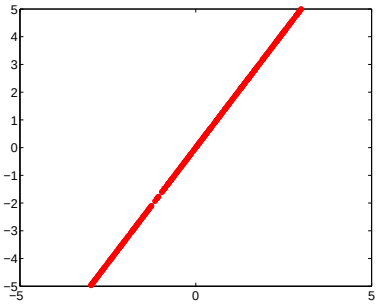
$r = 0$



$r = 0$

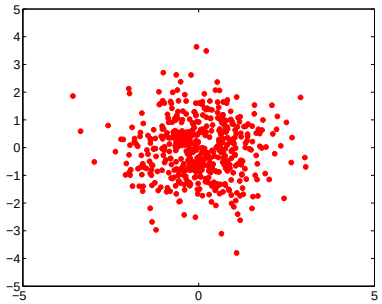


$r = 0$

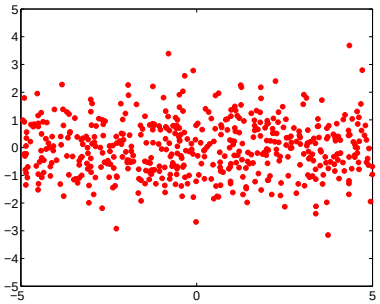


$r = 1$

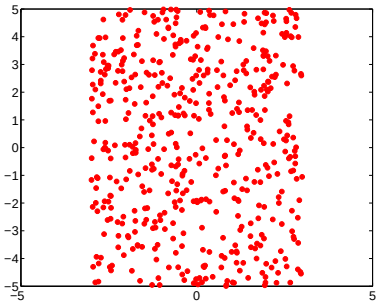
Korelace: příklady



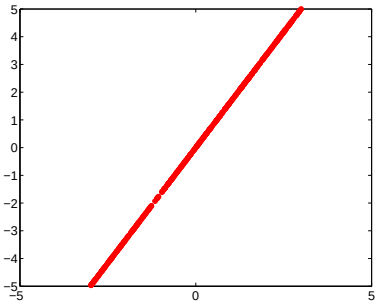
$r = 0$



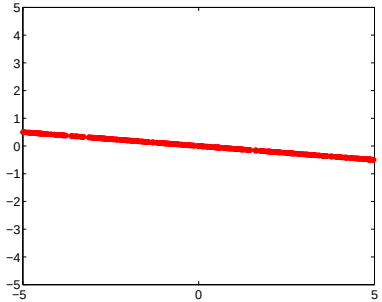
$r = 0$



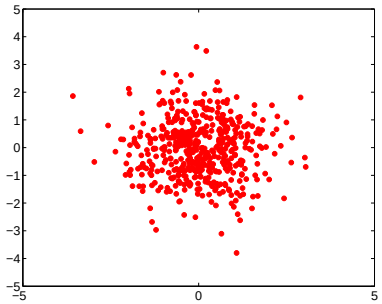
$r = 0$



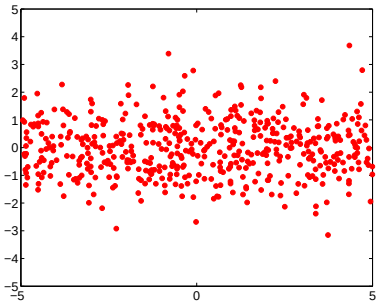
$r = 1$



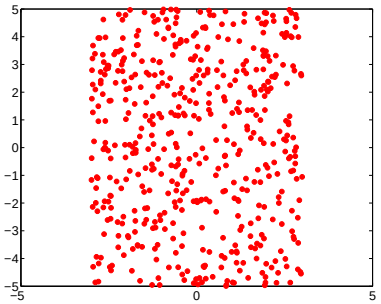
Korelace: příklady



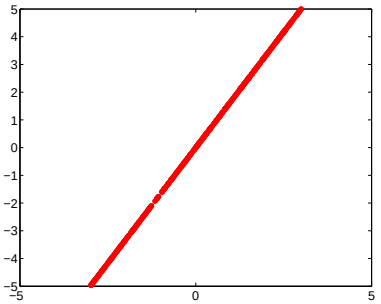
$r = 0$



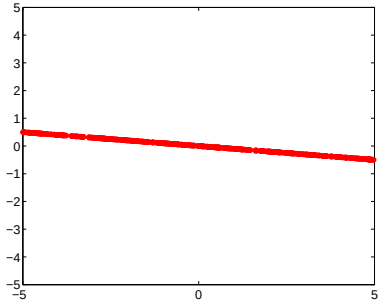
$r = 0$



$r = 0$

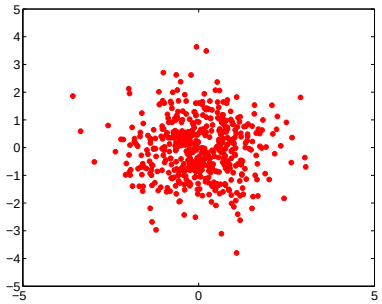


$r = 1$

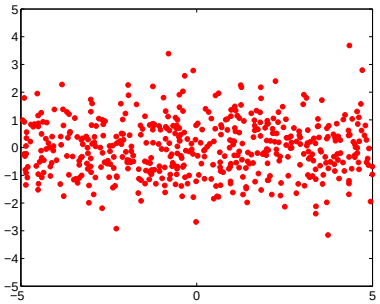


$r = -1$

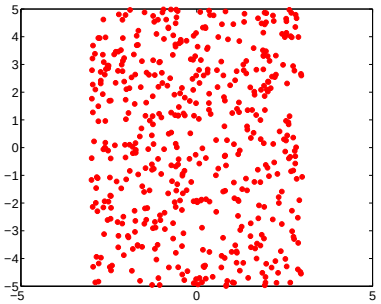
Korelace: příklady



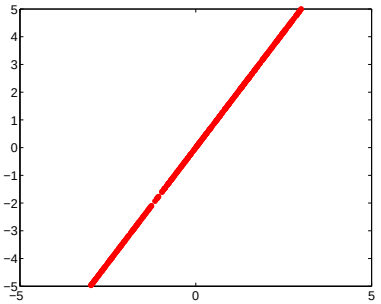
$r = 0$



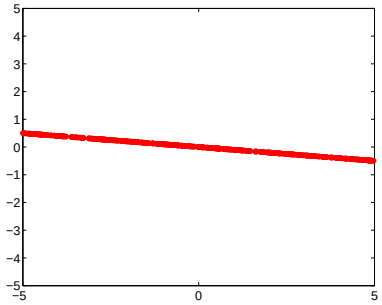
$r = 0$



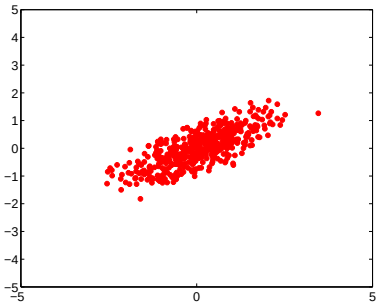
$r = 0$



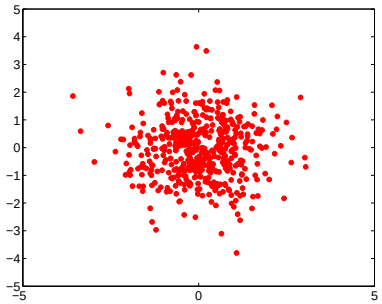
$r = 1$



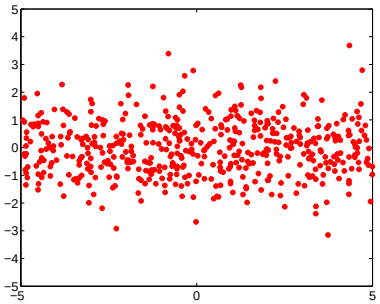
$r = -1$



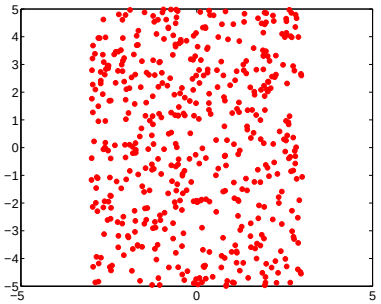
Korelace: příklady



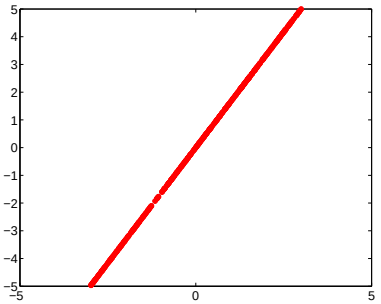
$r = 0$



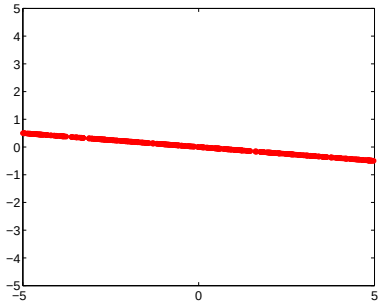
$r = 0$



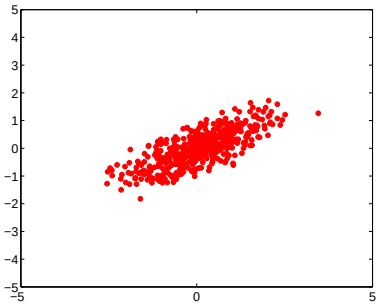
$r = 0$



$r = 1$

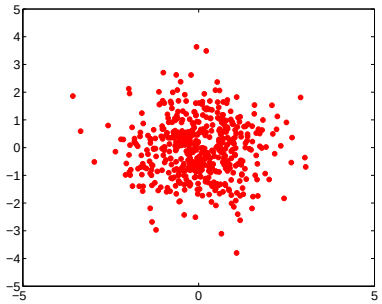


$r = -1$

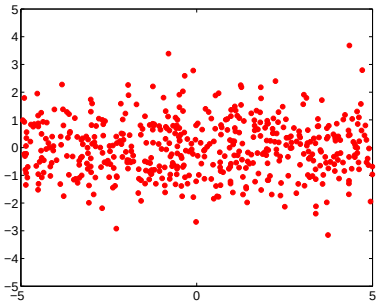


$r = 0.76$

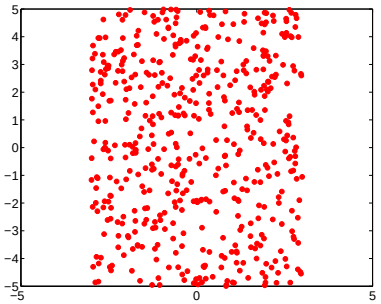
Korelace: příklady



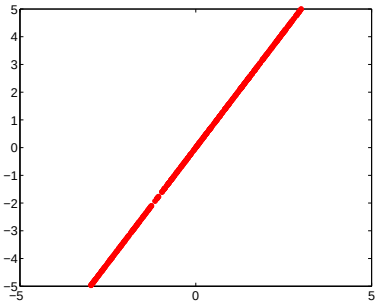
$r = 0$



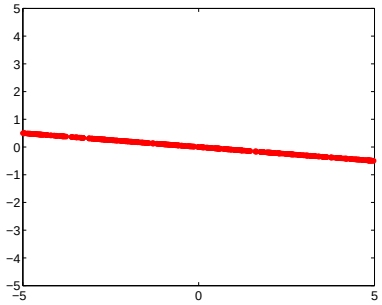
$r = 0$



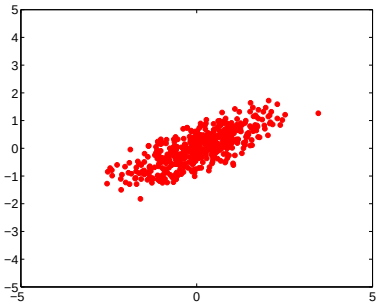
$r = 0$



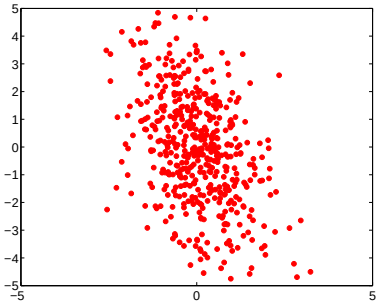
$r = 1$



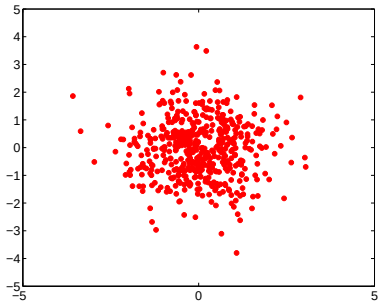
$r = -1$



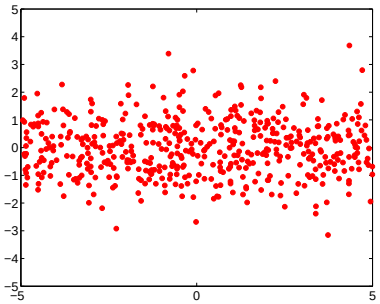
$r = 0.76$



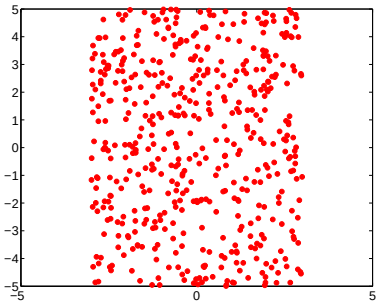
Korelace: příklady



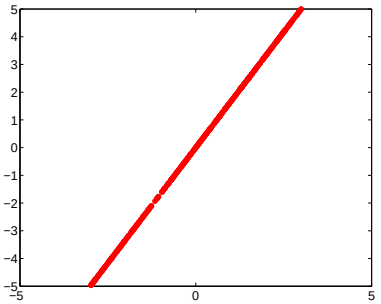
$r = 0$



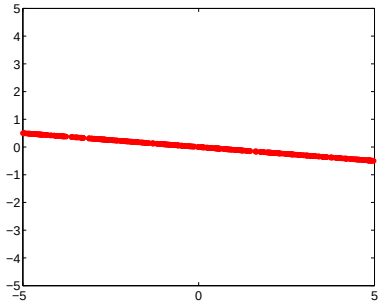
$r = 0$



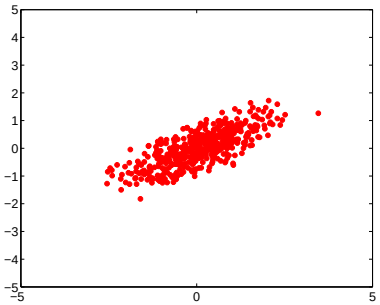
$r = 0$



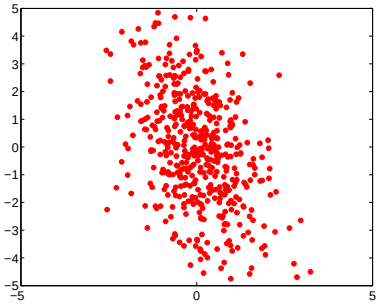
$r = 1$



$r = -1$

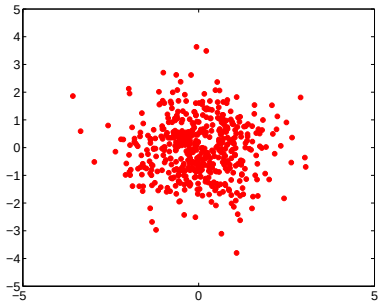


$r = 0.76$

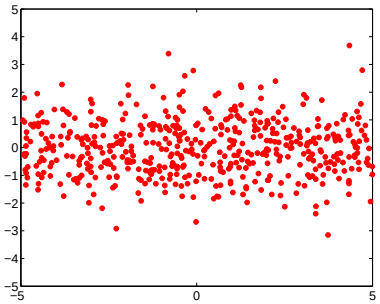


$r = -0.44$

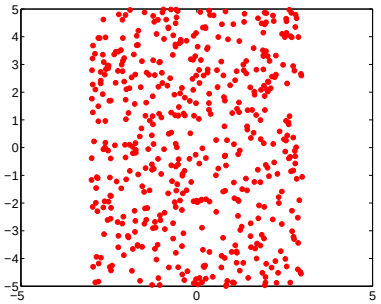
Korelace: příklady



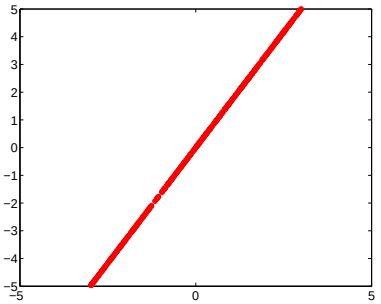
$r = 0$



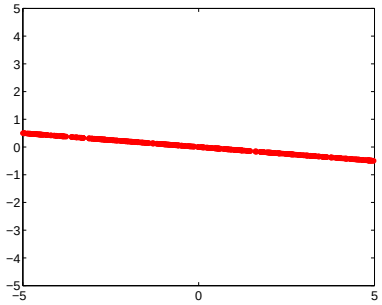
$r = 0$



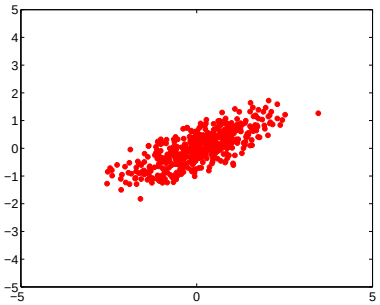
$r = 0$



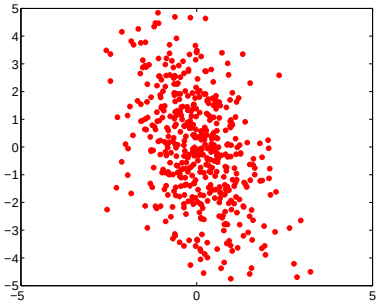
$r = 1$



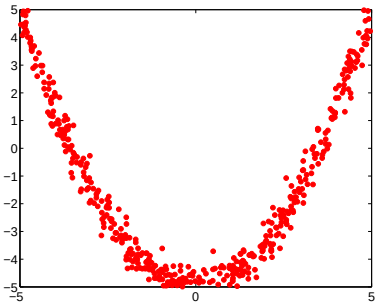
$r = -1$



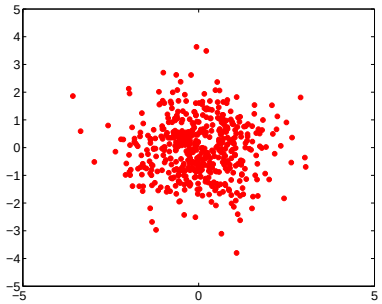
$r = 0.76$



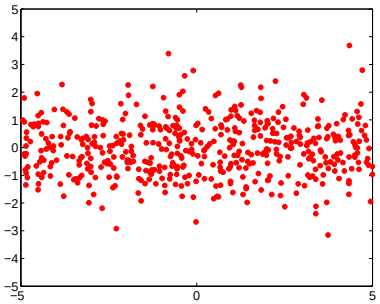
$r = -0.44$



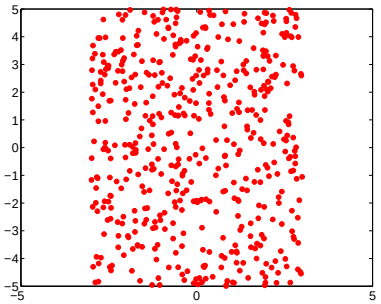
Korelace: příklady



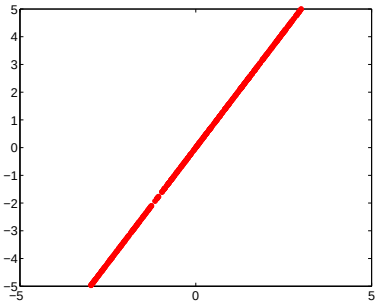
$r = 0$



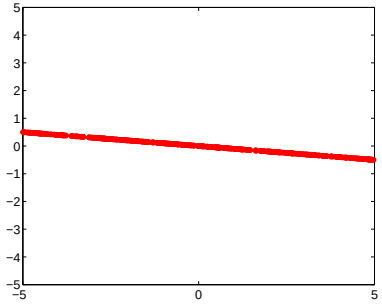
$r = 0$



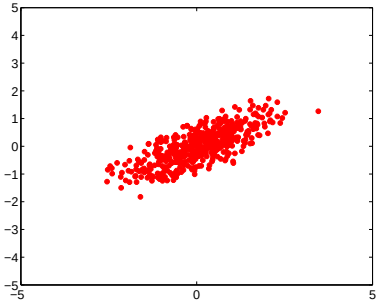
$r = 0$



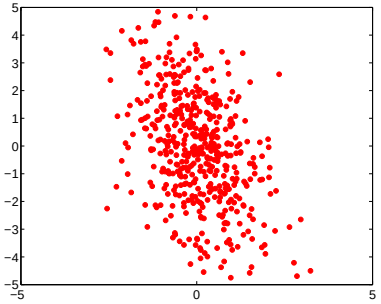
$r = 1$



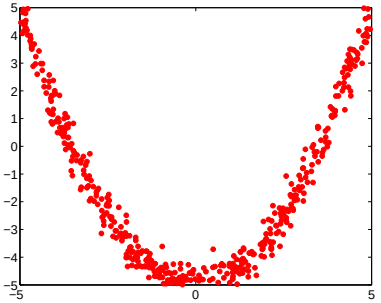
$r = -1$



$r = 0.76$



$r = -0.44$



$r = 0$

Korelace: shrnutí

- ✔ Pearsonův korelační koeficient měří sílu *lineární* závislosti
- ✔ Spearmanův nebo Kendallův pořadový korelační koeficient měří sílu *monotónní* závislosti
- ✔ Co dělat, když je závislost složitější (silně nelineární, nemonotonní)?
  - ✘ Diskretizace
  - ✘ 1D nelineární modely (viz další slidy)

Kritéria založená na úspěšnosti 1D modelů

Kritérium:

- ✔ Jak dobře lze modelovat výstup Y v závislosti na vstupu X_d?

Kritéria založená na úspěšnosti 1D modelů

Kritérium:

- ✔ Jak dobře lze modelovat výstup Y v závislosti na vstupu X_d?

Spojitá závislá veličina:

- ✔ Zvol typ nelineárního modelu f (např. regresní strom, polynomiální model)
- ✔ Spočti rozptyl závislé veličiny $V(Y)$ (popisuje neurčitost okolo průměru)
- ✔ Pro všechny vstupní proměnné X_d
  - ✘ Vytvoř model $\hat{Y} = f(X_d)$
  - ✘ Spočti rezidua modelu $R_d = Y - \hat{Y} = Y - f(X_d)$
  - ✘ Spočti rozptyl reziduí $V(R_d)$ (popisuje neurčitost okolo predikcí modelu)
- ✔ Čím nižší je $V(R_d)$, tím lepší model Y lze vytvořit na základě vstupu X_d
- ✔ Lze použít i jiné míry kvality regresních modelů (MSE, MAE, …)

Kritéria založená na úspěšnosti 1D modelů

Kritérium:

- ✔ Jak dobře lze modelovat výstup Y v závislosti na vstupu X_d?

Spojitá závislá veličina:

- ✔ Zvol typ nelineárního modelu f (např. regresní strom, polynomiální model)
- ✔ Spočti rozptyl závislé veličiny $V(Y)$ (popisuje neurčitost okolo průměru)
- ✔ Pro všechny vstupní proměnné X_d
  - ✘ Vytvoř model $\hat{Y} = f(X_d)$
  - ✘ Spočti rezidua modelu $R_d = Y - \hat{Y} = Y - f(X_d)$
  - ✘ Spočti rozptyl reziduí $V(R_d)$ (popisuje neurčitost okolo predikcí modelu)
- ✔ Čím nižší je $V(R_d)$, tím lepší model Y lze vytvořit na základě vstupu X_d
- ✔ Lze použít i jiné míry kvality regresních modelů (MSE, MAE, …)

Kategoriální závislá veličina:

- ✔ Postup obdobný
- ✔ Míry kvality klasifikačních modelů: chybovost, AUC (plocha pod ROC křivkou), …

Kritéria založená na teorii informace

Vzájemná informace:

- ✔ množství informace, kterou jedna náhodná veličina nese o druhé

$$I(X_d, Y) = \int_{x_d} \int_y p(x_d, y) \log \frac{p(x_d, y)}{p(x_d)p(y)} dx_d dy \tag{1}$$

- ✔ Teoreticky lze aplikovat na všechny typy proměnných
- ✔ V praxi se pro spojité X a Y obtížně odhaduje hustota pravděpodobnosti
- ✔ Pro kategoriální proměnné (nebo po diskretizaci spojitých)

$$I(X_d, Y) = \sum_{x_d} \sum_y P(X_d = x_d, Y = y) \log \frac{P(X_d = x_d, Y = y)}{P(X_d = x_d)P(Y = y)} \tag{2}$$

- ✔ Je-li vzájemná informace nulová, jsou veličiny nezávislé
- ✔ Čím větší vzájemná informace je, tím vyšší je souvislost mezi X_d a Y

Další statistická kritéria: chí-kvadrát test

χ^2 testy

- ✔ Manažer velké firmy na poradě vedení:
  Mám tu alarmující zprávu z poslední kontroly docházky našich zaměstnanců. Celých 40% sick-leavů připadá na pondělky a pátky! S tím musíme něco udělat!

Další statistická kritéria: chí-kvadrát test

χ^2 testy

- ✔ Manažer velké firmy na poradě vedení:
  Mám tu alarmující zprávu z poslední kontroly docházky našich zaměstnanců. Celých 40% sick-leavů připadá na pondělky a pátky! S tím musíme něco udělat!
- ✔ Titulek zprávy na zpravodajském portálu:
  Očkujte své děti! Celých 20 % onemocnění klíšťovou encefalitidou připadá na děti do 16 let.

Další statistická kritéria: chí-kvadrát test

χ^2 testy

- ✔ Manažer velké firmy na poradě vedení:
  Mám tu alarmující zprávu z poslední kontroly docházky našich zaměstnanců. Celých 40% sick-leavů připadá na pondělky a pátky! S tím musíme něco udělat!
- ✔ Titulek zprávy na zpravodajském portálu:
  Očkujte své děti! Celých 20 % onemocnění klíšťovou encefalitidou připadá na děti do 16 let.
- ✔ Velice populární a jednoduše pochopitelný test
- ✔ Díky malému množství předpokladů se obvykle řadí mezi testy *neparametrické*
- ✔ Měří se odchylka *pozorovaných (observed)* a *předpokládaných (expected)* četností v jednotlivých skupinách

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- ✔ Předpokládá se (H_0), že odchylky neexistují.
- ✔ Náhodná veličina χ^2 má rozdělení χ^2 s jistým počtem stupňů volnosti
- ✔ Velká hodnota χ^2 je důsledkem velké odchylky od H_0

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék		Celkem
	Streptomycin	Placebo	
Význ. zlepšení	28	4	32
Zlepšení	10	13	23
Beze změn	2	3	5
Zhoršení	5	12	17
Význ. zhoršení	6	6	12
Smrt	4	14	18
Celkem	55	52	107

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék		Celkem
	Streptomycin	Placebo	
Význ. zlepšení	28	4	32
Zlepšení	10	13	23
Beze změn	2	3	5
Zhoršení	5	12	17
Význ. zhoršení	6	6	12
Smrt	4	14	18
Celkem	55	52	107

$$E(\text{Výz. zlep.}, \text{Strept.}) = \frac{\text{Celkem(Výz. zlep.)} \cdot \text{Celkem(Strept.)}}{\text{Celkem}} =$$

$$= \frac{32 \cdot 55}{107} = 16.45$$

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék		Celkem
	Streptomycin	Placebo	
Význ. zlepšení	28	4	32
	16.45		
Zlepšení	10	13	23
Beze změn	2	3	5
Zhoršení	5	12	17
Význ. zhoršení	6	6	12
Smrt	4	14	18
Celkem	55	52	107

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék: Streptomycin	Lék: Placebo	Celkem
Význ. zlepšení	28	4	32
	16.45		
Zlepšení	10	13	23
Beze změn	2	3	5
Zhoršení	5	12	17
Význ. zhoršení	6	6	12
Smrt	4	14	18
Celkem	55	52	107

$$E(\text{Výz. zhor.}, \text{Placebo}) = \frac{\text{Celkem}(\text{Výz. zhor.}) \cdot \text{Celkem}(\text{Placebo})}{\text{Celkem}} =$$

$$= \frac{12 \cdot 52}{107} = 5.83$$

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék: Streptomycin	Lék: Placebo	Celkem
Význ. zlepšení	28 16.45	4	32
Zlepšení	10	13	23
Beze změn	2	3	5
Zhoršení	5	12	17
Význ. zhoršení	6	6 5.83	12
Smrt	4	14	18
Celkem	55	52	107

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék		Celkem
	Streptomycin	Placebo	
Význ. zlepšení	28	4	32
	16.45	15.55	
Zlepšení	10	13	23
	11.82	11.18	
Beze změn	2	3	5
	2.57	2.43	
Zhoršení	5	12	17
	8.74	8.26	
Význ. zhoršení	6	6	12
	6.17	5.83	
Smrt	4	14	18
	9.25	8.75	
Celkem	55	52	107

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék: Streptomycin	Lék: Placebo	Celkem
Význ. zlepšení	28 16.45	4 15.55	32
Zlepšení	10 11.82	13 11.18	23
Beze změn	2 2.57	3 2.43	5
Zhoršení	5 8.74	12 8.26	17
Význ. zhoršení	6 6.17	6 5.83	12
Smrt	4 9.25	14 8.75	18
Celkem	55	52	107

- ✔ Testová statistika se spočítá jako

$$\chi^2 = \sum_{i-1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 26.96$$

- ✔ Při r hodnotách 1. atributu a c hodnotách 2. atributu má taková náhodná veličina χ^2 rozdělení s $d.f. = (r-1)(c-1) = 5$ stupni volnosti
- ✔ Dosažaná hladina významnosti:

$$p = 1 - CDF\chi^2_5(26.96) = 5.8 \cdot 10^{-5}$$

Další statistická kritéria: Test nezávislosti

Zadání Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Řešení Použijeme χ^2-test nezávislosti:

Hodnocení	Lék		Celkem
	Streptomycin	Placebo	
Význ. zlepšení	28	4	32
	16.45	15.55	
Zlepšení	10	13	23
	11.82	11.18	
Beze změn	2	3	5
	2.57	2.43	
Zhoršení	5	12	17
	8.74	8.26	
Význ. zhoršení	6	6	12
	6.17	5.83	
Smrt	4	14	18
	9.25	8.75	
Celkem	55	52	107

- ✔ Testová statistika se spočítá jako

$$\chi^2 = \sum_{i-1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 26.96$$

- ✔ Při r hodnotách 1. atributu a c hodnotách 2. atributu má taková náhodná veličina χ^2 rozdělení s $d.f. = (r-1)(c-1) = 5$ stupni volnosti
- ✔ Dosažaná hladina významnosti:

$$p = 1 - CDF\chi^2_5(26.96) = 5.8 \cdot 10^{-5}$$

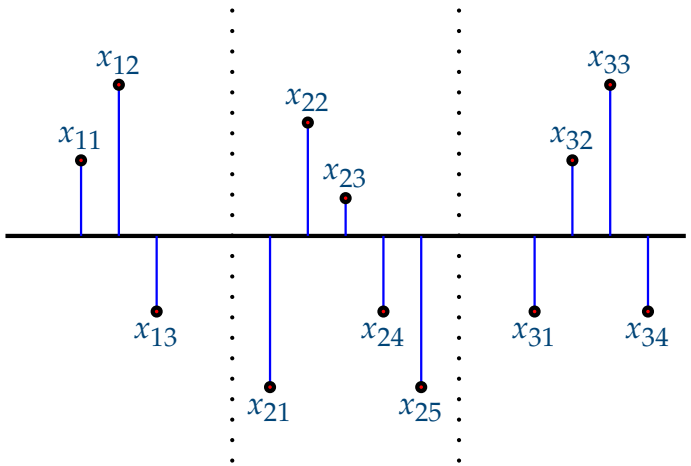
Celková ochylka od předpokládaných hodnot (29.96) je natolik velká, že bychom ji za předpokladu nezávislosti obou atributů mohli pozorovat jen s velmi malou pravděpodobností ($5.8 \cdot 10^{-5}$). Je mnohem pravděpodobnější, že předpoklad nezávislosti neplatí.

Další statistická kritéria: Analýza rozptylu

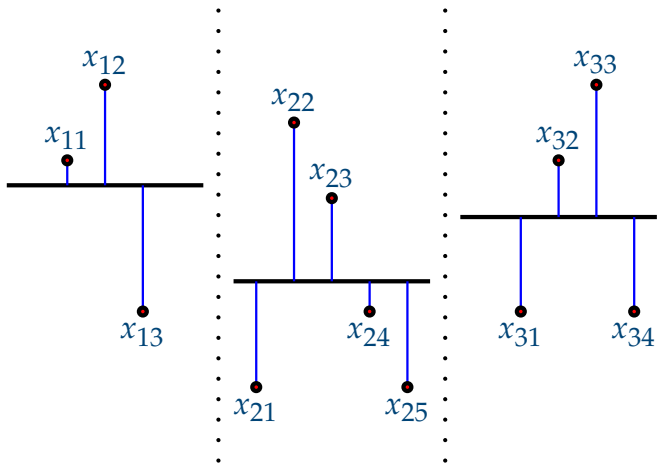
Jednofaktorová analýza rozptylu (ANOVA)

- ✔ Posouzení souvislosti mezi 1 nominální a 1 spojitou proměnnou.
- ✔ Data kategorizujeme podle nominální proměnné.
  - ✘ Máme několik skupin dat, mají rozdělení $N(\mu_i, \sigma^2)$
- ✔ Chceme zjistit, zda hodnota nominální proměnné ovlivňuje očekávanou hodnotu spojité proměnné.
  - ✘ Předpokládáme nezávislost, t.j. $H_0 : \mu_1 = \mu_2 = ... = \mu_a$.
  - ✘ Počítáme odchylku od předpokládaného stavu: $SS_A = SS_T - SS_E$
  - ✘ Čím je odchylka větší, tím výraznější souvislost mezi veličinami existuje.

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2$$



$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$



ANOVA: Předpoklady

- ✔ Nezávislost jednotlivých pozorování
- ✔ Nezávislost jednotlivých skupin
- ✔ Normální rozdělení sledované veličiny ve všech skupinách
  Kolmogorov-Smirnovův test, Shapiro-Wilkův test, χ^2 test dobré shody
- ✔ Shoda rozptylů ve skupinách
  Bartlettův test, Levenův test, Hartleyův test
- ✔ Při porušení posledních dvou předpokladů je možné použít neparametrickou, tzv. Kruskal-Wallisovu ANOVu.

Proč 1D metody nestačí?

Nadbytečné (redundantní) proměnné?

Nadbytečná proměnná

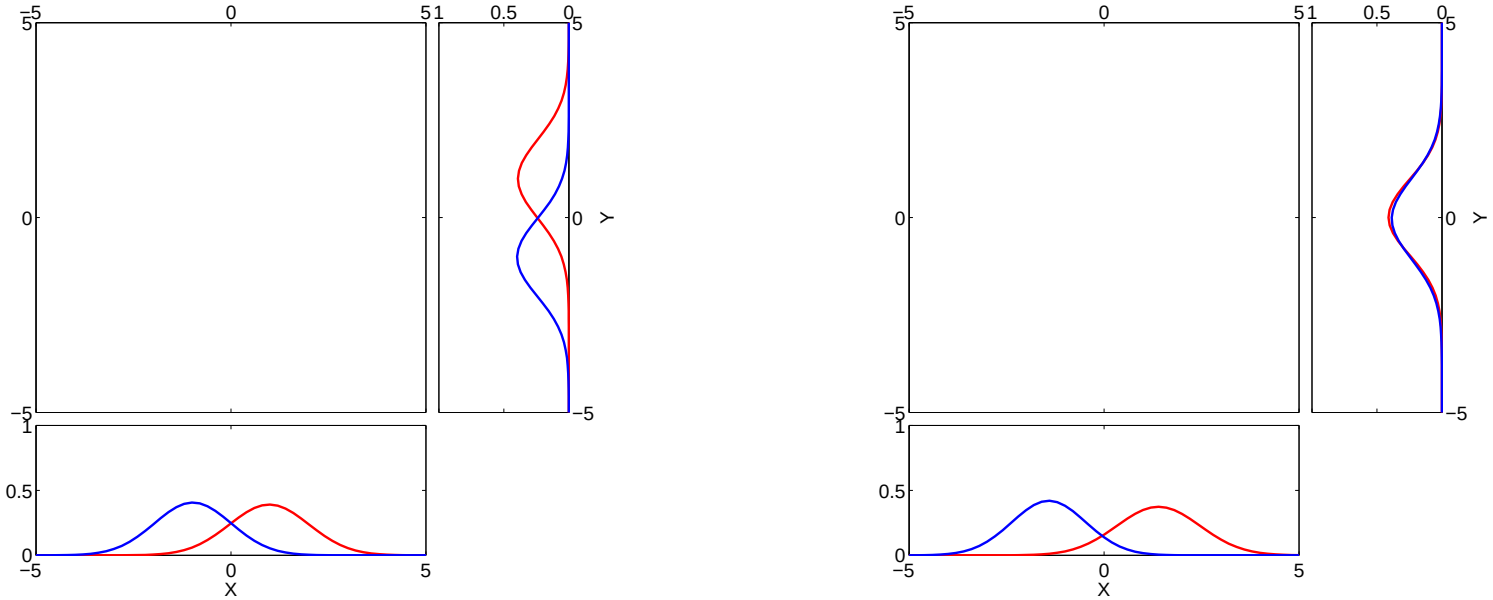
- ✔ nese o závislé proměnné stejnou informaci, jako jiná vstupní proměnná

Nadbytečné (redundantní) proměnné?

Nadbytečná proměnná

- ✔ nese o závislé proměnné stejnou informaci, jako jiná vstupní proměnná

Lze nadbytečnost proměnné posoudit z 1D průmětů?



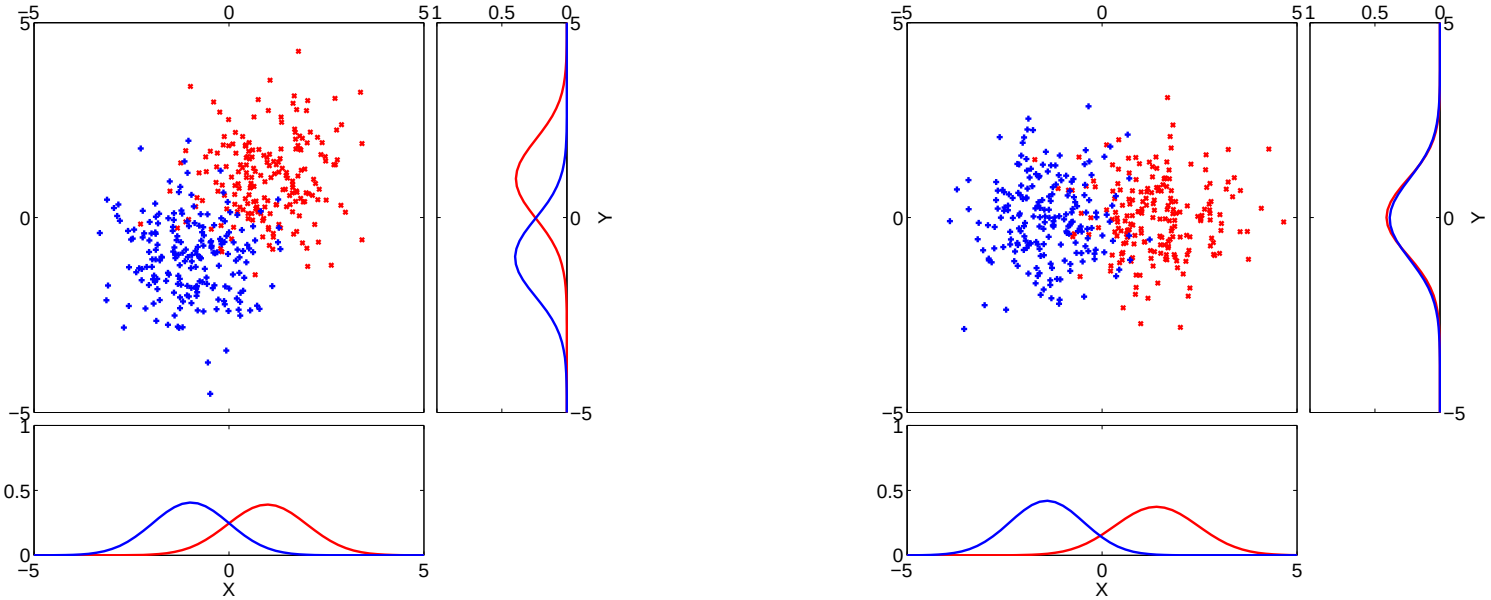
- ✔ Z 1D průmětů se zdá, že obě proměnné vlevo mají přibližně stejnou vypovídací schopnost a že jedna z nich je tedy redundantní. Vpravo je jedna proměnná zdánlivě zbytečná (Y), druhá (X) se zdá být užitečnější než obě proměnné vlevo (“kopce” jsou dále od sebe)

Nadbytečné (redundantní) proměnné?

Nadbytečná proměnná

- ✔ nese o závislé proměnné stejnou informaci, jako jiná vstupní proměnná

Lze nadbytečnost proměnné posoudit z 1D průmětů?



- ✔ Z 1D průmětů se zdá, že obě proměnné vlevo mají přibližně stejnou vypovídací schopnost a že jedna z nich je tedy redundantní. Vpravo je jedna proměnná zdánlivě zbytečná (Y), druhá (X) se zdá být užitečnější než obě proměnné vlevo (“kopce” jsou dále od sebe)
- ✔ Data vpravo jsou ale jen pootočenou verzí dat vlevo. Kdybychom jednu z proměnných vlevo vyřadili, nebyli bychom schopni vytvořit situaci vpravo.

Vliv korelace na redundanci?

Předchozí slide: uvnitř tříd žádné korelace nebyly, ale

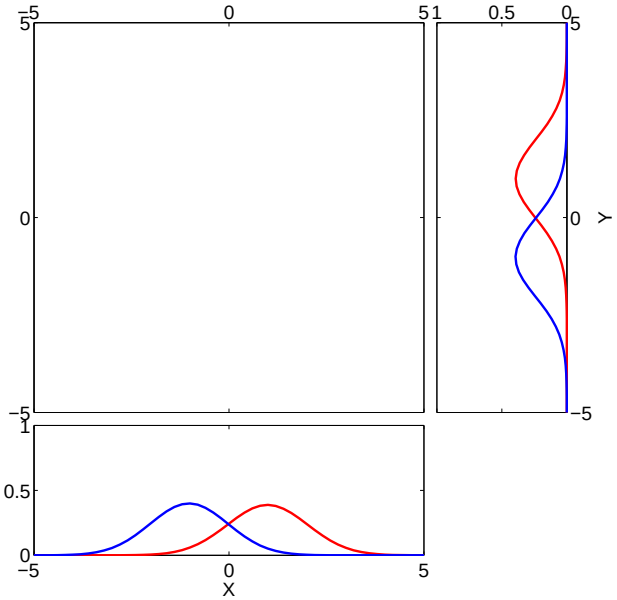
- ✔ proměnné byly částečně korelované díky posunu středů Gaussiánů.

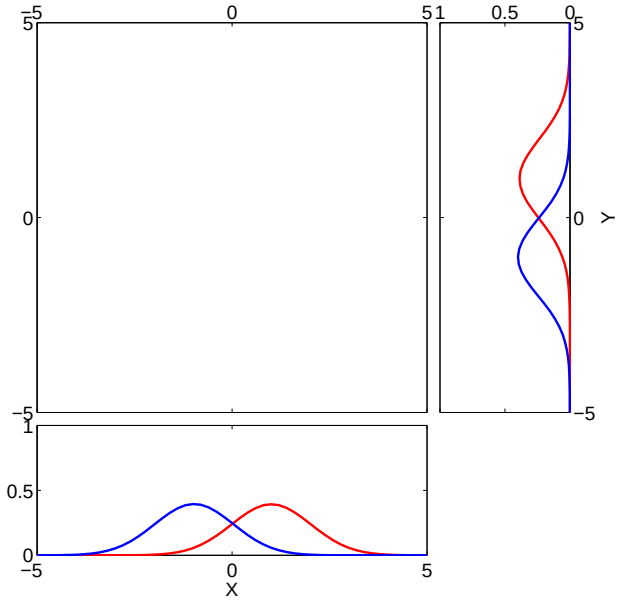
Vliv korelace na redundanci?

Předchozí slide: uvnitř tříd žádné korelace nebyly, ale

- ✔ proměnné byly částečně korelované díky posunu středů Gaussiánů.

Jak ovlivňuje korelace redundanci?





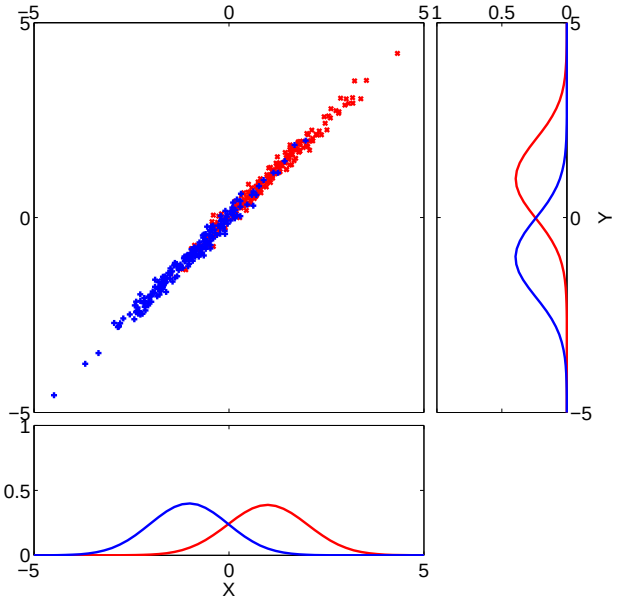
- ✔ 1D průměty do proměnných X a Y vlevo i vpravo vypadají naprosto stejně.

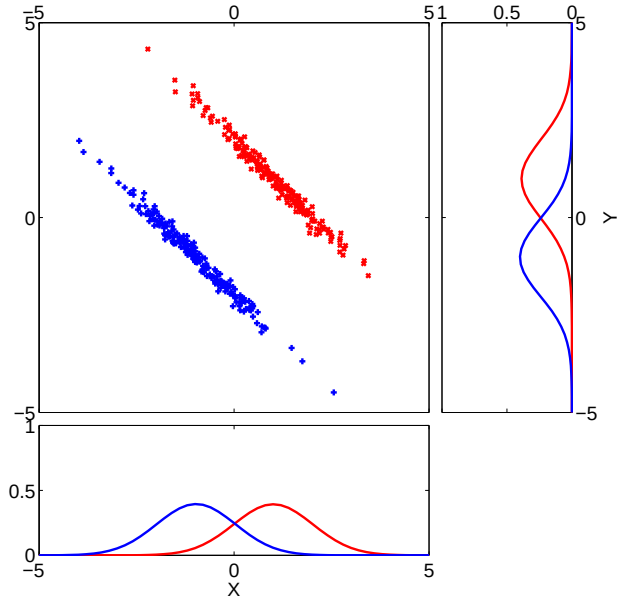
Vliv korelace na redundanci?

Předchozí slide: uvnitř tříd žádné korelace nebyly, ale

- ✔ proměnné byly částečně korelované díky posunu středů Gaussiánů.

Jak ovlivňuje korelace redundanci?





- ✔ 1D průměty do proměnných X a Y vlevo i vpravo vypadají naprosto stejně.
- ✔ Vlevo jsou proměnné velmi korelované, jedna je téměř lineární funkcí druhé, jedna proměnná je zde skutečně redundantní — nese naprosto stejnou informaci jako druhá. Vpravo je situace jiná: obě třídy jsou zcela jasně separovatelné; kdybychom jednu z proměnných vyřadili, nebylo by možné třídy separovat.

Zbytečné proměnné?

Zbytečná proměnná

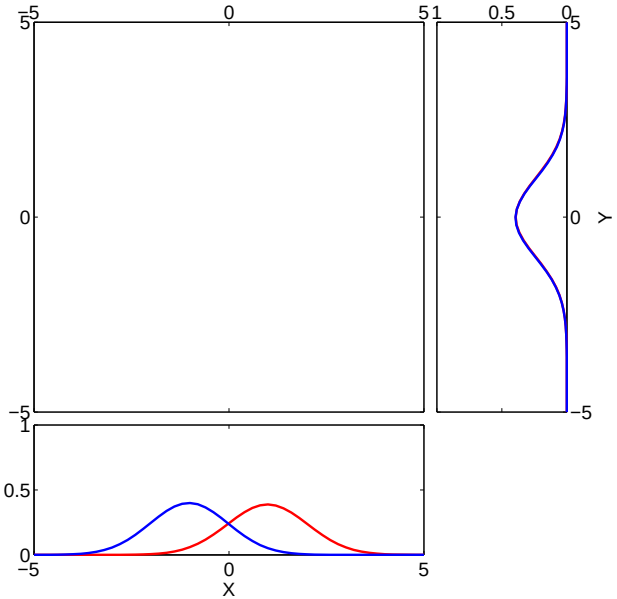
- ✔ nenese o závislé proměnné žádnou informaci, výstup je na ní nezávislý

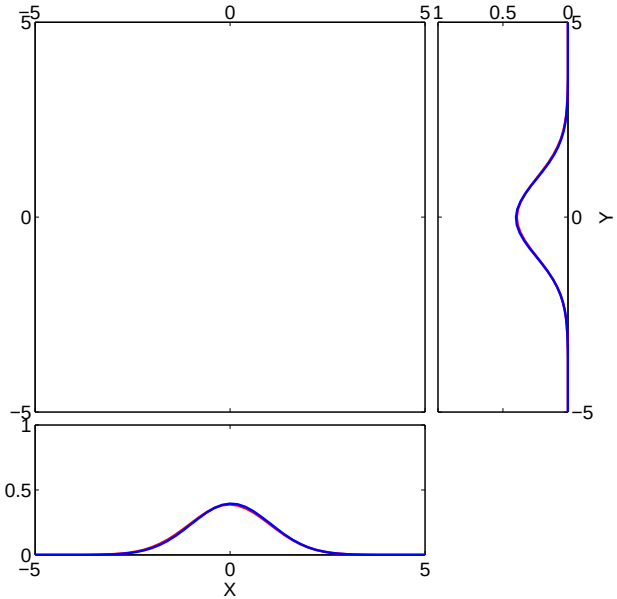
Zbytečné proměnné?

Zbytečná proměnná

- ✔ nenese o závislé proměnné žádnou informaci, výstup je na ní nezávislý

Lze zbytečnost proměnné posoudit z 1D průmětů? Může být zdánlivě zbytečná proměnná užitečná v kombinaci s jinými?





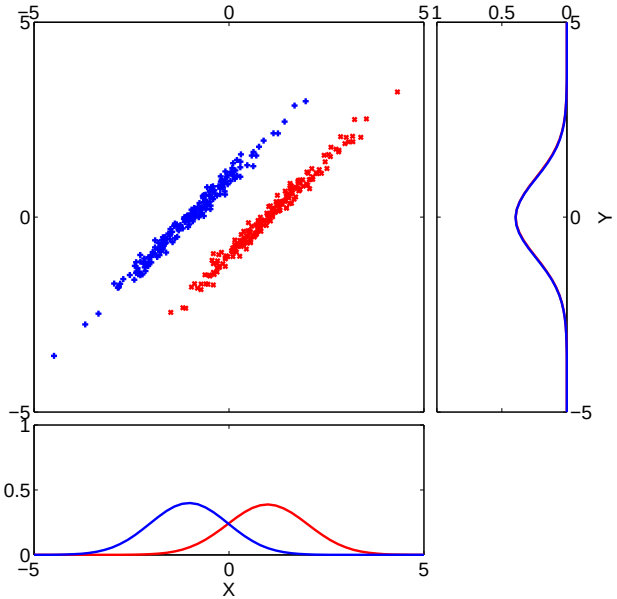
- ✔ Vlevo: Z 1D průmětů se zdá, že proměnná X informaci o třídě nese, zatímco proměnná Y nikoli. Vpravo: informaci o třídě nenese zdánlivě ani jedna proměnná.

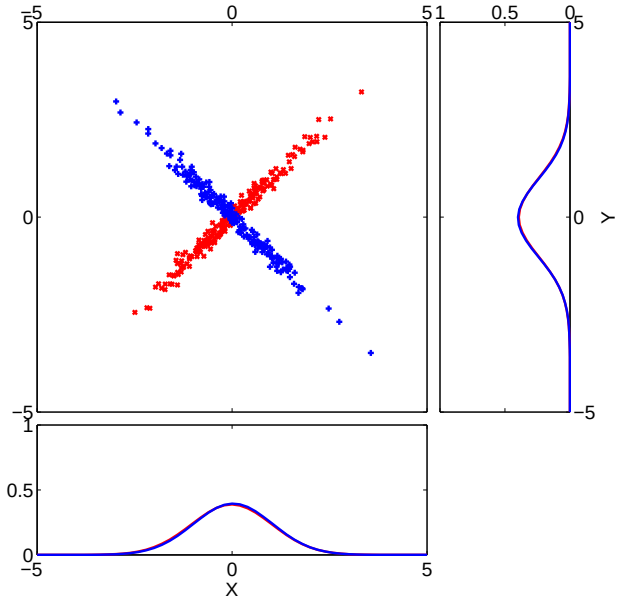
Zbytečné proměnné?

Zbytečná proměnná

- ✔ nenese o závislé proměnné žádnou informaci, výstup je na ní nezávislý

Lze zbytečnost proměnné posoudit z 1D průmětů? Může být zdánlivě zbytečná proměnná užitečná v kombinaci s jinými?





- ✔ Vlevo: Z 1D průmětů se zdá, že proměnná X informaci o třídě nese, zatímco proměnná Y nikoli. Vpravo: informaci o třídě nenese zdánlivě ani jedna proměnná.
- ✔ Vlevo: zdánlivě zbytečná proměnná Y je užitečná v kombinaci s X! Vpravo: ačkoli jsou obě proměnné zdánlivě zbytečné, obě dohromady umožní sestavit poměrně dobrý klasifikátor! Vyřazením zdánlivě zbytečných proměnných můžeme udělat velkou chybu!

Mnoharozměrné metody výběru proměnných

Mnoharozměrné metody výběru proměnných

Jednorozměrné metody mohou selhat:

- ✔ Nepoznají, že proměnná vliv má (v kombinaci s jinou proměnnou).
- ✔ Jako významné určí skupinu proměnných, které jsou mezi sebou závislé, t.j. stačí zařadit do modelu jen 1 z nich.

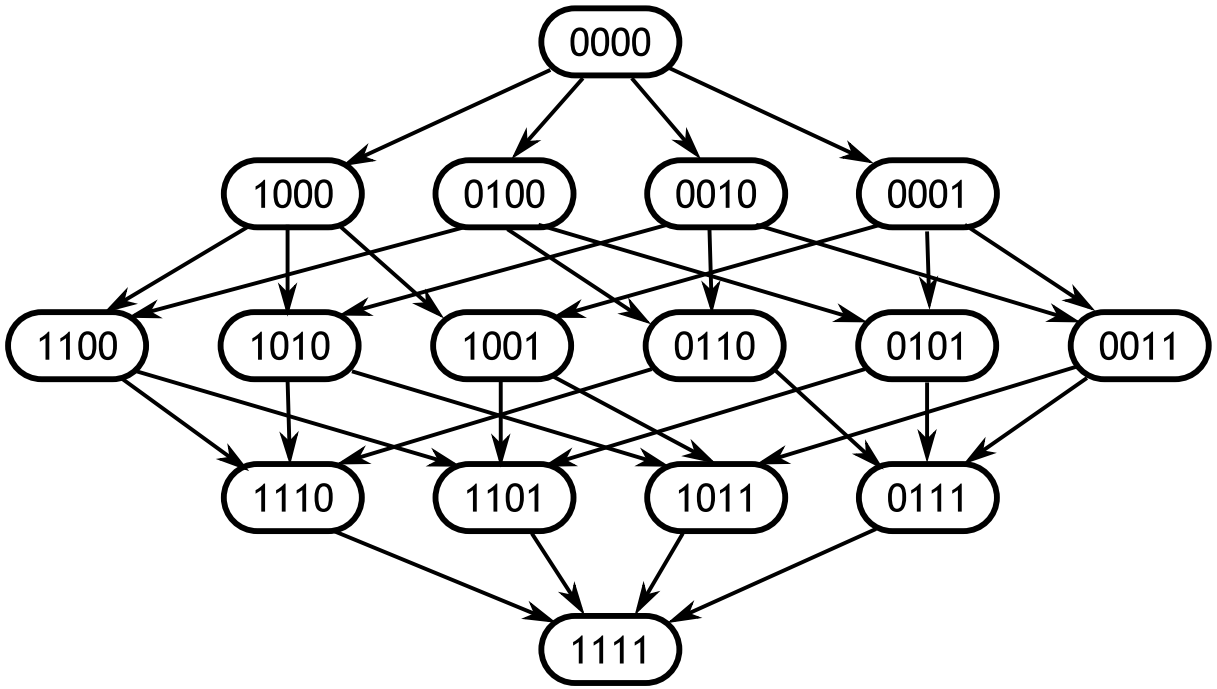
Mnoharozměrné metody výběru proměnných

Jednorozměrné metody mohou selhat:

- ✔ Nepoznají, že proměnná vliv má (v kombinaci s jinou proměnnou).
- ✔ Jako významné určí skupinu proměnných, které jsou mezi sebou závislé, t.j. stačí zařadit do modelu jen 1 z nich.

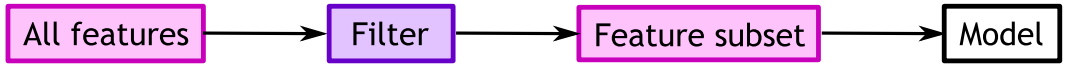
Mnoharozměrný výběr proměnných je složitý!

- ✔ N proměnných, 2^N různých podmnožin!



Filter vs. Wrapper

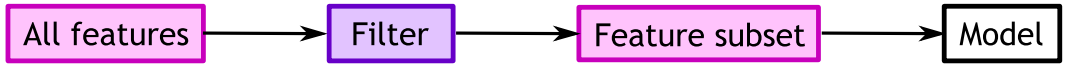
Filter: vybírá podskupinu proměnných nezávisle na modelu



- ✔ jednorázový proces
- ✔ poskytne sadu "nejvýznamnějších" proměnných jako výslednou vybranou podmnožinu *nezávisle na použitém modelu*

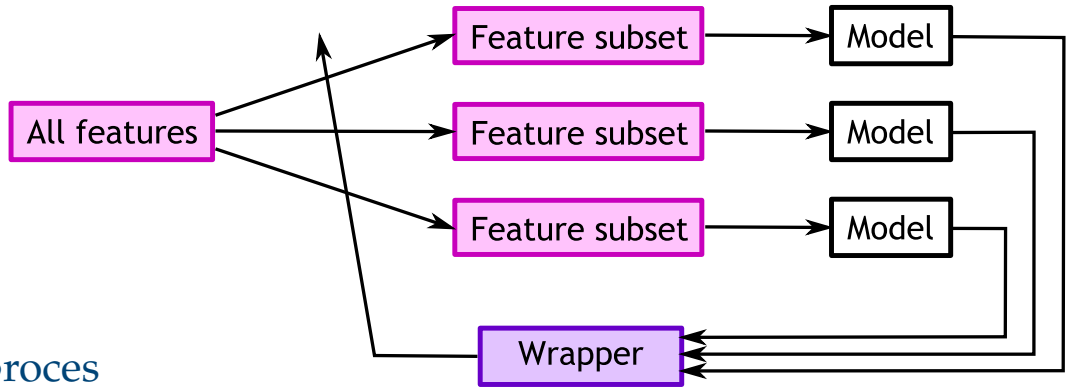
Filter vs. Wrapper

Filter: vybírá podskupinu proměnných nezávisle na modelu



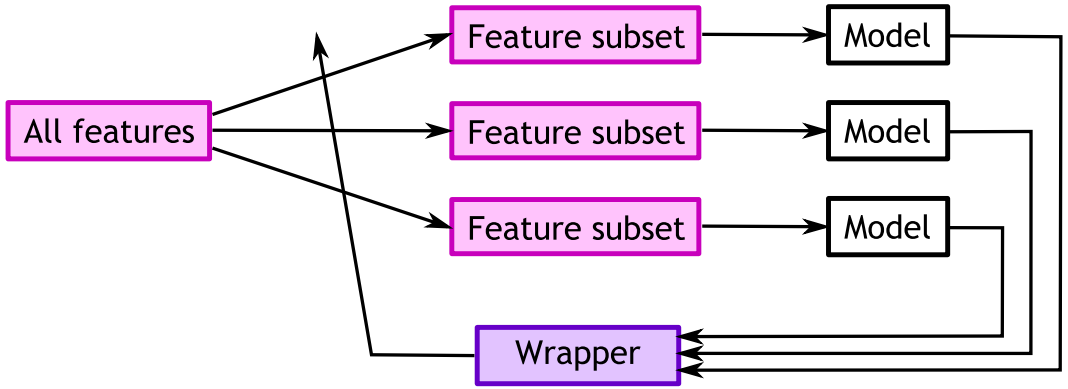
- ✔ jednorázový proces
- ✔ poskytne sadu "nejvýznamnějších" proměnných jako výslednou vybranou podmnožinu *nezávisle na použitém modelu*

Wrapper: vybírá podskupinu proměnných s ohledem na model



- ✔ iterativní proces
- ✔ v každé iteraci vygeneruje několik podmnožin vst. proměnných a otestuje jejich přínos na *konkrétním typu modelu*
- ✔ podle úspěšnosti modelu na jednotlivých podmnožinách aktivně ovlivňuje podmnožiny proměnných vybrané k testování v další iteraci

Wrappers



Wrapper

- ✔ Zcela obecná metoda selekce proměnných
- ✔ Typ modelu a jeho algoritmus učení jsou černou skříňkou, nijak se nemění

Před použitím je třeba definovat:

- ✔ Jaký typ modelu a jaký učicí algoritmus bude použit?
- ✔ Jak hodnotit přesnost modelu? (vede hledání, určuje, kdy se hledání zastaví)
  - ✘ Testovací data nebo krosvalidace
- ✔ Jak prohledávat prostor možných podmnožin proměnných?
  - ✘ NP-těžký problém
  - ✘ Enumerativní prohledávání možné při malém počtu proměnných (viz kosatce v úvodu)
  - ✘ Hladové prohledávání (*forward selection* nebo *backward elimination*)
  - ✘ Metoda větví a mezí, simulované žíhání, genetické algoritmy, ...

Extrakce proměnných

Extrakce proměnných

Lepší výsledky se často dosahují pomocí proměnných odvozených z původních vstupů (agregace, transformace, . . .):

- ✔ Vypovídá o změně pacientova stavu spíše to, že byl na kontrole v září 2008, říjnu 2009, v lednu 2010, v únoru 2010 a v dubnu 2010, nebo to, že minulé dva roky byl na kontrole vždy jen 1x, zatímco v první třetině roku 2010 už 3x?
- ✔ Bude pro odhadování výsledku šachové partie lepší, když budete vědět, že černý král je na D1 a bílá dáma na H4, nebo to, když víte, že bílá dáme ohrožuje černého krále?

Odvozují (konstruují) se nové proměnné, které mohou být

- ✔ lineární i nelineární funkcí
- ✔ jedné, více, nebo všech vstupních proměnných a které by měly mít
- ✔ větší souvislost s modelovanou závislou proměnnou.
- ✔ Často se využívá doménových znalostí.

Dva různé cíle metod extrakce proměnných:

- ✔ co nejlepší rekonstrukce dat
- ✔ co největší užitečnost pro predikce

Metody:

- ✔ shlukování (skupina “podobných” proměnných je nahrazena jejich centroidem)
- ✔ analýza hlavních komponent (PCA/SVD), projection pursuit, lineární diskriminační analýza (LDA), kernel PCA, . . .
- ✔ spektrální transformace (Fourierova, vlnková), . . .

Závěr

Shrnutí

- ✔ Selekce optimální podmnožiny vstupních proměnných je NP-těžký problém.
- ✔ 1D metody jsou
  - ✘ jednoduché a umožňují seřadit proměnné podle zdánlivé užitečnosti pro predikci,
  - ✘ v praxi poměrně často fungují, ale
  - ✘ mohou se dopustit fatálních chyb.
- ✔ Mnoharozměrné metody jsou
  - ✘ odolnější proti chybám při selekci, ale
  - ✘ výpočetně mnohem náročnější.
- ✔ Rozlišujeme
  - ✘ filtry,
  - ✘ wrappery a
  - ✘ embedded metody.
- ✔ Selekce proměnných pouze vybírá podmonžinu vstupních proměnných, extrakce konstruuje zcela nové proměnné na základě původních.

Reference

[GE03] Isabelle Guyon and Andre Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.