Pattern Recognition

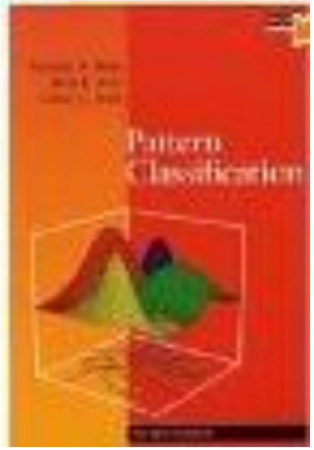
Lectures: J. Matas and V. Franc (today, and when JM has other commitments)

About the Lectures

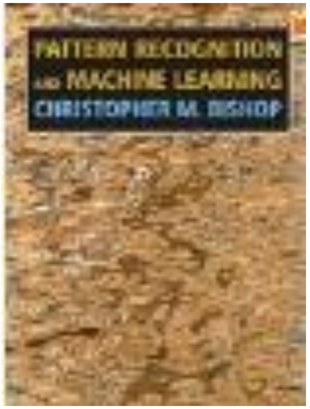
- The selection of the topics in the course is mainstream. Besides course material, a good wiki page is available for almost all topic cover in the course.
- We strongly recommend attendance of lectures. In PR&ML, many issues are intertwined and it is very difficult to understand the connections (e.g. understanding "why method X should be used instead of Y in case Z") just by reading about particular methods.
- Nevertheless, we do not introduce any "incentives" e.g. in the form of a written exam during a lecture.

- No single textbooks is ideal for Pattern Recognition and Machine Learning course. The field is still waiting for one ....

Textbooks



- **Duda, Hart, Stork: Pattern Classification**
  classical text, 2nd edition, “easy reading”,
  about 5-10 available at CMP library (G102, R. Kopecka will lend you a copy)
  some sections obsolete



- **Bishop: Pattern Recognition and Machine Learning**
  new, popular, but certain topics, in my opinion, could be presented in a clearer way

- **Schlesinger, Hlavac: Ten Lectures on Statistical and Structural Pattern Recog.**
  advanced text, for those who want to know more than what is presented in the course;
  aims at maximum generality

English/Czech Lectures

- Those of you who are fulfilling the requirement of OI to chose one course in English should attend the lecture in English, i.e. on Monday. It is acceptable to attend the Friday lectures a few times if you miss the one on Monday.

- You may attend *both* lectures (a couple of students did this last year to gain better understanding). Note that after the 28.10, the Czech version will run one week late.

- If English terminology is unclear, ask. As most of the terms will be used repeatedly, language problems will disappear over time.

Pattern Recognition

The course focuses on *statistical pattern recognition.*

We start with an example called “Dilemma of a lazy short-sighted student of OI” which introduces most of the basic ingredients of a statistical decision problem.

Example: A lazy short-sighted OI student dilemma.

A student with a weak eyesight and a strong dislike for running is in a hurry. He needs to get to Albertov, where he has arranged to play a poker game. He might get there on time, but he needs to catch a tram immediately. The club rules stipulate he’ll have to pay 100 CZK fine if he’s late.

As he exits Building A at Karlovo namesti, he sees a tram at the stop. He cannot see the tram number as he is short-sighted, but he recognises the tram is the rectangular shaped “new style” one, not the rounded “old style”.

Should he run?

Example: A lazy short-sighted student dilemma …

The student prefers well-justified, and if possible, optimal decisions. He travels to Albertov regularly and he knows:

- #18 and #24 go to Albertov
- the following trams stop at Karlovo namesti: 3,6,18,22,24
- the joint probability P(x,k) a tram of type x ∈ {old, new} and number k ∈ {3,6,18,22,24} is:

P(x,k)	3	6	18	22	24	P(k)
old type	0.05	0.15	0. 10	0.25	0.05	0.60
new type	0.20	0. 00	0. 05	0.00	0.15	0.40
P(k)	**0.25**	**0. 15**	**0. 15**	**0.25**	**0.20**	

So should he run?

Example: A lazy short-sighted student dilemma …

P(x,k)	3	6	18	22	24	P(k)
old type	0.05	0.15	0. 10	0.25	0.05	0.60
new type	0.20	0. 00	0. 05	0.00	0.15	0.40

P(k)	0.25	0. 15	0. 15	0.25	0.20

The notation P(x,k), P(x), P(k) is a shorthand that can lead to ambiguities. Here the meaning of P(old) and P(3) is clear.

But if trams were of type 1, 2 and 3, P(3) would be ambiguous. In that case, the notation $P(\boldsymbol{X}{=}x)$, $P(\boldsymbol{X}{=}x, \boldsymbol{K}{=}k)$ will be used, e.g. $P(\boldsymbol{X}{=}\text{old}, \boldsymbol{K}{=}18)$. Some textbooks use a $P_{XK}(x,k)$, $P_K(k)$ notation; $P_K(3)$ is the probability $P(\boldsymbol{K}{=}3)$.

The probabilities P(x) and P(k) are called marginal.
In pattern recognition literature, P(k) is called *a priori probability*.

Example: A lazy short-sighted student dilemma …

P(x,k)	3	6	18	22	24	P(k)
old type	0.05	0.15	0. 10	0.25	0.05	0.60
new type	0.20	0. 00	0. 05	0.00	0.15	0.40

P(k)	0.25	0. 15	0. 15	0.25	0.20

The notation P(x,k), P(x), P(k) is a shorthand that can lead to ambiguities. Here the meaning of P(old) and P(3) is clear.

But if trams were of type 1, 2 and 3, P(3) would be ambiguous. In that case, the notation $P(\boldsymbol{X}{=}x)$, $P(\boldsymbol{X}{=}x,\boldsymbol{K}{=}k)$ will be used, e.g. $P(\boldsymbol{X}{=}\text{old},\ \boldsymbol{K}{=}18)$. Some textbooks use a $P_{XK}(x,k)$, $P_K(k)$ notation; $P_K(3)$ is the probability $P(\boldsymbol{K}{=}3)$.

The probabilities P(x) and P(k) are called marginal.
In pattern recognition literature, P(k) is called *a priori probability*.

Example: A lazy short-sighted student dilemma …

P(x,k)	3	6	18	22	24	P(k)
old type	0.05	0.15	0. 10	0.25	0.05	0.60
new type	0.20	0. 00	0. 05	0.00	0.15	0.40
P(k)	**0.25**	**0. 15**	**0. 15**	**0.25**	**0.20**	

The notation P(x,k), P(x), P(k) is a shorthand that can lead to ambiguities. Here the meaning of P(old) and P(3) is clear.

But if trams were of type 1, 2 and 3, P(3) would be ambiguous. In that case, the notation $P(X{=}x)$, $P(X{=}x,K{=}k)$ will be used, e.g. $P(X{=}\text{old},\ K{=}18)$. Some textbooks use a $P_{XK}(x,k)$, $P_K(k)$ notation; $P_K(3)$ is the probability $P(K{=}3)$.

The probabilities P(x) and P(k) are called marginal.
In pattern recognition literature, P(k) is called *a priori probability*.

OI Student Dilemma

So should he run?

We still do not know enough to give a well-justified advice.

We know that by missing a tram #18 or 24 he'll loose 100 CZK.

Our student is money driven. In his life, everything can be converted to financial loss or gain. He values a needless run to be a loss of 50 CZK.

The advice will have a form of a strategy. In this example, thre are only four strategies possible:

1. if you see an old tram, run, else don't run (and miss it)
2. if you see a new tram, run, else don't run
3. never run
4. always run

Q: what is the number of strategies in the general case with D possible decisions and |X| observations ?

OI Student Dilemma

So should he run?

Indeed, we now have enough information to find the optimal strategy for the short-sighted, lazy, money driven OI student.

Exact mathematical formulation of the problem and how to find the optimal solution is exactly what you'll will learn in Lecture #2.

The solution of this particular problem will be presented next week.

A lazy short-sighted OI student dilemma, v.2

Let's consider the following modification:

A student with a weak eyesight and a strong dislike for running is in a hurry. He needs to get to Albertov, where his girlfriend, a medical student is expecting him in 10 minutes. He might get there on time, but he needs to catch a tram immediately.

As he exits Building A at Karlovo namesti, he sees a tram at the stop. He cannot see the tram number as he is short-sighted, but he recognizes the tram is the rectangular shaped "new style" one, not the rounded "old style".

He knows, as before, X,K, P(X,K).

He knows his girlfriend tolerates him being late 20% of the time and does not even comment. But she'd dump him if gets aboven that.

When should he run?

A lazy short-sighted OI student dilemma, v.2

This problem will be studied in lecture #3.

Interestingly, in this case, the student need not assign a cost to running or to the loosing his girlfriend (which might be rather difficult).

He needs a strategy that will tell him to run as rarely as possible, given the constraint: he must catch the tram 80% of time else he looses his girlfriend.

Formulation of the Statistical PR Problem

In a stat. PR problem, the following is given:

- X the set of observations (measurements)
- K the set of "hidden states" (or "classes"). The state cannot be directly observed
- D the set of decisions
- P(X,K) the probabilistic model of the relationship between the observations and the hidden state
- (for the so called Bayesian problem) W: K x D $\rightarrow$ R the loss function. W(k,d) gives the loss incurred by taking a decision d $\in$ D if the object is in the (unknown) state k $\in$ K

The solution of the problem is a function q: X $\rightarrow$ D, called a strategy, that for defines a decision for every observation.

The quality of the strategy q can s measured by a number of ways, the expected (average) loss is the most common.

Classification

Very often, the sets of states K and decisions D coincide. Such problem is called **classification.**

Example: In a drink vending machine, classify coins according to their value. The set of measurements could be say weight, diameter and electrical resistance, i.e. $X = R^3$. The set of hidden classes is $K =$ {1,2,5,10,20,50}, the set of decisions D=K.

Note: in many places the designer of the machine will soon discover the need to enlarge the set of decision D by a “not a cot coin” class.

Examples of statistical PR problems

Application	X	D
finger printer verification	an image	grant/refuse access
optical character recognition	an image	non-character, a-zA-Z0_9…
speaker identification	sound recording	1 of N of known identities
banknote check	different sensors	genuine, forgery
ECG check	m recordings of voltage over time	healthy / diagnosis A, B, ..
vending machine	some measurement on coins	value of the coin (0 if not a coin)
dictation machine	sound recording	word
spam filter	email content, sender info, …	spam / no spam
automatic check reading	an image	amount on the check

Examples of statistical PR problems: Notes

- For many examples, most possible observation x will never appear, for most no x will be observed more than once.
- For most of the listed examples, there is therefore no hope of knowing P(X,K)
- For some of the examples, estimate the cardinality of the space of observations X.
- For some of the examples, estimate the cardinality of the space of all possible strategies Q.

The Statistical PR Problem. Notes.

The formulation given is very general. As seen in the example, the cardinalities of X and D (K) range from 2 to infinite.

For many applications, the formulation captures all important aspects. Nevertheless, other important aspect were ignored, e.g.:

- The choice of X, which was assumed given. In many applications, the choice of X, is left for the designer.
- The cost and time of making a measurement was ignored. With a cheap camera, observations arrive instantly and at minimum cost (of powering the camera). In medical applications, each measurement is costly (disposable material like vials, expensive hardware to take a scan, labor costs)
- The time to decision, a strategy was characterized only by its loss.
- The measurements x were viewed as inputs. In many decision processes, e.g. seeing a doctor, values of initial measurements define what measurements will be made next.

The Statistical PR Problem. Notes.

- In some problems, the hidden state k cannot be observed *in principle.* Example: K is the value of the dollar against CZK tomorrow. X is the exchange rate for each day in the last year. Decisions are “sell USD now” or “buy USD now”.

- Often the “hidden state” is potentially observable, but at a large cost. It is practical to equip notebooks with fingerprint readers and solve a statistical PR problem, with acceptable precision. A DNA analyzer would be error-free, but too costly.

Statistical Decision Problems not Covered

Another assumption made in the PR problem formulation (and in the “Lazy Student Dilemma”) is that the object does not react to our measurements or decision.

For instance, a nasty tram driver might close the door faster when he sees the lazy student running. Another driver might wait as soon as he sees the students starts to run, and he catches the tram just walking fast. Such situations are outside the pattern recognition domain, they are studied in **game theory.**

In statistical PR problems covered in the course, a single decision (classification) is made, time does not enter the formulation. **Control theory** studies feedback systems, where observations (measurements), decision (actions) and time play a critical role, infinite number of decisions are made

Statistical Decision Problems not Covered

Syntactic analysis. Consider the C compiler problem. Given a finite text input, it must make a binary decision:

syntactically valid/invalid C program.

The problem is complex, but not statistical,

The statistical model P(X,K) and the Training set

- The first part of the course is about solving statistical pattern recognition problems when the model P(X,K) is known.

- It is very rare that P(X,K) is known for a given application. Instead, it is almost always possible to obtain a set of representative samples T of (measurement, class) pairs, Example: Gender recognition. A person labels 1000 face images man / woman.

- One way to proceed is to find and estimate $P'(X,Y)$ from T and proceed as if the estimate was equal to the true probability.

- A much more common approach is to obtain a strategy q (= a classifier) with desirable properties directly from T.

macros_rpz.tex
sfmath.sty
cmpitemize.tex

Thank you for your attention.