

Probability estimation

Tomáš Svoboda

thanks to Ondřej Drbohlav, Michal Reinstein, Jiří Matas

Vision for Robots and Autonomous Systems, Center for Machine Perception

Department of Cybernetics

Faculty of Electrical Engineering, Czech Technical University in Prague

May 22, 2019

Shortened presentation

- ▶ This is a shortened version of the presentation.
- ▶ It does not cover a gentle introduction into MLE principle.
- ▶ It assumes students know about Maximum likelihood estimation from a Probability and Statistics course.

Probability estimation

For simplicity we assume 1-dim (scalar) features x as far we can

In previous two lectures:

$$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}}$$

In practice:

- ▶ unknown quantities
- ▶ estimate from training data $\mathcal{T} = \{(x_1, s_1), (x_2, s_2), \dots, (x_I, s_I)\}$

Problem: tossing coing, is it fair, how is the $P(\text{head})$?



Does Max Likelihood (ML) solve it all?

- ▶ Tossing coin, $\mathcal{T} = \{T, T, T\}$
- ▶ What the ML estimate of p_H ?
- ▶ Would you believe it?
- ▶ What is missing?

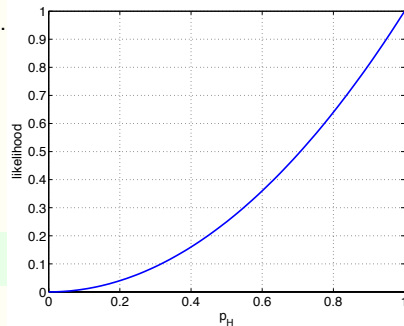
Tossed 2 \times , two heads $\mathcal{T} = \{H, H\}$.
We assume iid.

$$P(H, H | p_H = 0.5) = 0.5^2 = 0.25$$

$$P(H, H | p_H = 0.2) = 0.2^2 = 0.04$$

$$P(H, H | p_H = 0.8) = 0.8^2 = 0.64$$

Likelihood: $\mathcal{L}(p_H | \mathcal{T})$



$x_n = 1$ if H, and $x_n = 0$ for T.

$$\mathcal{L}(p_H | \mathcal{T}) = p(\mathcal{T} | p_H) = \prod_{i=1}^N p(x_n | p_H) = \prod_{i=1}^N p_H^{x_n} (1 - p_H)^{1 - x_n}$$

Bernoulli distribution. What is the best p_H ?

Log the whole product and ∂p_H , and at the end, ...

$$p_H = \frac{\sum x_n}{N}$$

Bernoulli distribution is a special case of Binomial distribution for $n = 1$.

Does Max Likelihood (ML) solve it all?

- ▶ Tossing coin, $\mathcal{T} = \{T, T, T\}$
- ▶ What the ML estimate of p_H ?
- ▶ Would you believe it?
- ▶ What is missing?

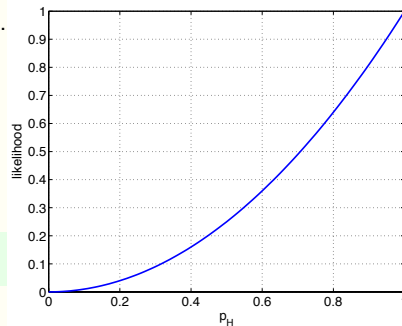
Tossed 2x, two heads $\mathcal{T} = \{H, H\}$.
We assume iid.

$$P(H, H | p_H = 0.5) = 0.5^2 = 0.25$$

$$P(H, H | p_H = 0.2) = 0.2^2 = 0.04$$

$$P(H, H | p_H = 0.8) = 0.8^2 = 0.64$$

Likelihood: $\mathcal{L}(p_H | \mathcal{T})$



$x_n = 1$ if H, and $x_n = 0$ for T.

$$\mathcal{L}(p_H | \mathcal{T}) = p(\mathcal{T} | p_H) = \prod_{i=1}^N p(x_i | p_H) = \prod_{i=1}^N p_H^{x_i} (1 - p_H)^{1 - x_i}$$

Bernoulli distribution. What is the best p_H ?

Log the whole product and ∂p_H , and at the end, ...

$$p_H = \frac{\sum x_n}{N}$$

Bernoulli distribution is a special case of Binomial distribution for $n = 1$.

Does Max Likelihood (ML) solve it all?

- ▶ Tossing coin, $\mathcal{T} = \{T, T, T\}$
- ▶ What the ML estimate of p_H ?
- ▶ Would you believe it?
- ▶ What is missing?

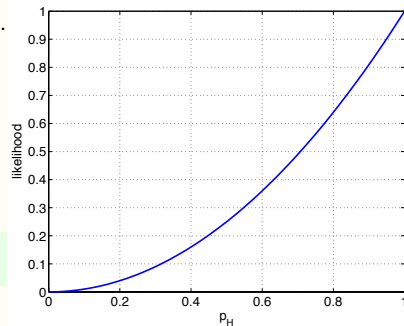
Tossed 2 \times , two heads $\mathcal{T} = \{H, H\}$.
We assume iid.

$$P(H, H | p_H = 0.5) = 0.5^2 = 0.25$$

$$P(H, H | p_H = 0.2) = 0.2^2 = 0.04$$

$$P(H, H | p_H = 0.8) = 0.8^2 = 0.64$$

Likelihood: $\mathcal{L}(p_H | \mathcal{T})$



$x_n = 1$ if H, and $x_n = 0$ for T.

$$\mathcal{L}(p_H | \mathcal{T}) = p(\mathcal{T} | p_H) = \prod_{i=1}^N p(x_n | p_H) = \prod_{i=1}^N p_H^{x_n} (1 - p_H)^{1 - x_n}$$

Bernoulli distribution. What is the best p_H ?

Log the whole product and ∂p_H , and at the end, ...

$$p_H = \frac{\sum x_n}{N}$$

Bernoulli distribution is a special case of Binomial distribution for $n = 1$.

Tossing coin, using priors

$$\mathcal{L}(p_H|\mathcal{T}) = p(\mathcal{T}|p_H) = \prod_{i=1}^N p(x_n|p_H) = \prod_{i=1}^N p_H^{x_n} (1 - p_H)^{1-x_n}$$

$$p(h, N|p_H) = \binom{N}{h} p_H^h (1 - p_H)^{N-h}; \quad p_H = \frac{h}{N}$$

(Conjugate) Prior:

$$p(p_H|a, b) \sim p_H^a (1 - p_H)^b$$

Conjugate because the likelihood and the prior have the same form.

The prior $p(p_H|a, b)$ is actually the Beta distribution, https://en.wikipedia.org/wiki/Beta_distribution

Tossing coin, using priors

$$\mathcal{L}(p_H|\mathcal{T}) = p(\mathcal{T}|p_H) = \prod_{i=1}^N p(x_n|p_H) = \prod_{i=1}^N p_H^{x_n} (1 - p_H)^{1-x_n}$$

$$p(h, N|p_H) = \binom{N}{h} p_H^h (1 - p_H)^{N-h}; \quad p_H = \frac{h}{N}$$

(Conjugate) Prior:

$$p(p_H|a, b) \sim p_H^a (1 - p_H)^b$$

Conjugate because the likelihood and the prior have the same form.

The prior $p(p_H|a, b)$ is actually the Beta distribution, https://en.wikipedia.org/wiki/Beta_distribution

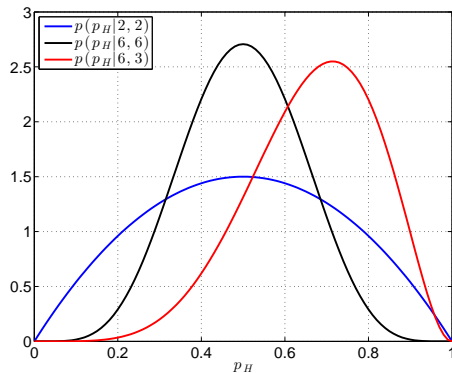
Tossing coin, using priors

$$\mathcal{L}(p_H|\mathcal{T}) = p(\mathcal{T}|p_H) = \prod_{i=1}^N p(x_n|p_H) = \prod_{i=1}^N p_H^{x_n} (1-p_H)^{1-x_n}$$

$$p(h, N|p_H) = \binom{N}{h} p_H^h (1-p_H)^{N-h}; \quad p_H = \frac{h}{N}$$

(Conjugate) Prior:

$$p(p_H|a, b) \sim p_H^a (1-p_H)^b$$



Conjugate because the likelihood and the prior have the same form.

The prior $p(p_H|a, b)$ is actually the Beta distribution, https://en.wikipedia.org/wiki/Beta_distribution

Using the prior

$$p(h, N | p_H) \sim p_H^h (1 - p_H)^{N-h}$$

$$p(p_H | a, b) \sim p_H^a (1 - p_H)^b$$

$$p(p_H | h, N) \sim p(h, N | p_H) p(p_H) \sim p_H^{h+a} (1 - p_H)^{N-h+b}$$

Looking for extremum

$$\frac{\partial p(p_H | h, N)}{\partial p_H} = 0$$

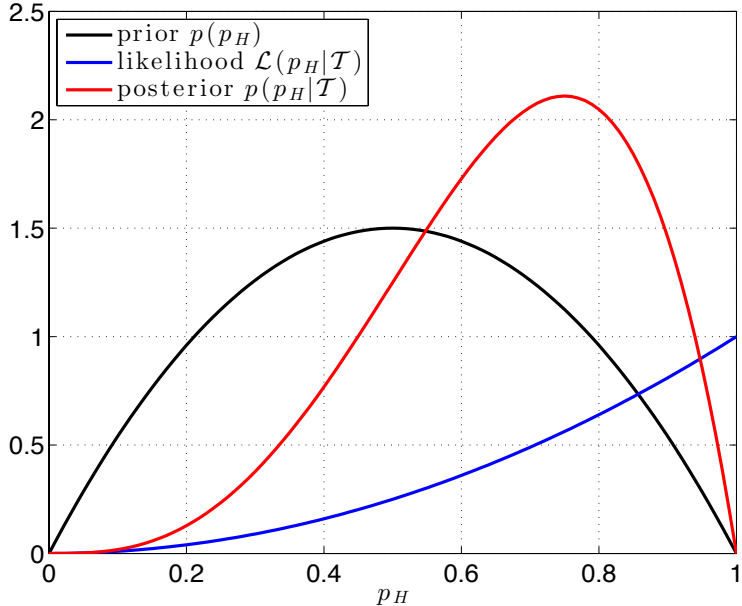
yields

$$p_H = \frac{h + a}{N + a + b}$$

Hyperparameters a, b as regularization

Maximum a posteriori estimate

See the map.m demo.



Problem: Coins classification based on weight

s/x	5 g	10 g	15 g	20 g	25 g	Σ
1 CZK	15	10	3	0	0	28
2 CZK	7	13	16	6	1	43
5 CZK	0	1	2	11	15	29
Σ	22	24	21	17	16	100

- Two weighting devices A, B . $x_A = 16$, $x_B = 19$ what to do?

Maximum likelihood estimation of the weight

Two weighting devices A, B with some σ_A, σ_B measure $x_A = 16, x_B = 19$.

What is the ML estimate of the weight w ?

► Devices independent:

$$\mathcal{L}(w) = p(x_A, x_B | w) = p(x_A | w)p(x_B | w)$$

► Sensors Gaussian:

$$\mathcal{L}(w) = \frac{1}{\sigma_A \sqrt{2\pi}} \exp\left[-\frac{(x_A - w)^2}{2\sigma_A^2}\right] \times \frac{1}{\sigma_B \sqrt{2\pi}} \exp\left[-\frac{(x_B - w)^2}{2\sigma_B^2}\right]$$

$$\ell(w) = \ln \dots$$

after some derivation, ..., weighted average

$$w = \frac{x_A \sigma_A^{-2} + x_B \sigma_B^{-2}}{\sigma_A^{-2} + \sigma_B^{-2}}$$

Why is Gaussian assumption reasonable, it is at all? Central Limit Theorem, see `clt.m` demo.

Maximum likelihood estimation of the weight

Two weighting devices A, B with some σ_A, σ_B measure $x_A = 16, x_B = 19$.
What is the ML estimate of the weight w ?

► Devices independent:

$$\mathcal{L}(w) = p(x_A, x_B | w) = p(x_A | w)p(x_B | w)$$

► Sensors Gaussian:

$$\mathcal{L}(w) = \frac{1}{\sigma_A \sqrt{2\pi}} \exp\left[-\frac{(x_A - w)^2}{2\sigma_A^2}\right] \times \frac{1}{\sigma_B \sqrt{2\pi}} \exp\left[-\frac{(x_B - w)^2}{2\sigma_B^2}\right]$$

$$\ell(w) = \ln \dots$$

after some derivation, ..., weighted average

$$w = \frac{x_A \sigma_A^{-2} + x_B \sigma_B^{-2}}{\sigma_A^{-2} + \sigma_B^{-2}}$$

Why is Gaussian assumption reasonable, it is at all? Central Limit Theorem, see `clt.m` demo.

Maximum likelihood estimation of the weight

Two weighting devices A, B with some σ_A, σ_B measure $x_A = 16, x_B = 19$.
What is the ML estimate of the weight w ?

- ▶ Devices independent:

$$\mathcal{L}(w) = p(x_A, x_B | w) = p(x_A | w)p(x_B | w)$$

- ▶ Sensors Gaussian:

$$\mathcal{L}(w) = \frac{1}{\sigma_A \sqrt{2\pi}} \exp\left[-\frac{(x_A - w)^2}{2\sigma_A^2}\right] \times \frac{1}{\sigma_B \sqrt{2\pi}} \exp\left[-\frac{(x_B - w)^2}{2\sigma_B^2}\right]$$

$$\ell(w) = \ln \dots$$

after some derivation, ..., weighted average

$$w = \frac{x_A \sigma_A^{-2} + x_B \sigma_B^{-2}}{\sigma_A^{-2} + \sigma_B^{-2}}$$

Why is Gaussian assumption reasonable, it is at all? Central Limit Theorem, see `clt.m` demo.

Maximum likelihood estimation of the weight

Two weighting devices A, B with some σ_A, σ_B measure $x_A = 16, x_B = 19$.
What is the ML estimate of the weight w ?

- ▶ Devices independent:

$$\mathcal{L}(w) = p(x_A, x_B | w) = p(x_A | w)p(x_B | w)$$

- ▶ Sensors Gaussian:

$$\mathcal{L}(w) = \frac{1}{\sigma_A \sqrt{2\pi}} \exp \left[-\frac{(x_A - w)^2}{2\sigma_A^2} \right] \times \frac{1}{\sigma_B \sqrt{2\pi}} \exp \left[-\frac{(x_B - w)^2}{2\sigma_B^2} \right]$$

$$\ell(w) = \ln \dots$$

after some derivation, ..., weighted average

$$w = \frac{x_A \sigma_A^{-2} + x_B \sigma_B^{-2}}{\sigma_A^{-2} + \sigma_B^{-2}}$$

Why is Gaussian assumption reasonable, it is at all? Central Limit Theorem, see `clt.m` demo.

Problem: Coins classification based on weight

s/x	5 g	10 g	15 g	20 g	25 g	Σ
1 CZK	15	10	3	0	0	28
2 CZK	7	13	16	6	1	43
5 CZK	0	1	2	11	15	29
Σ	22	24	21	17	16	100

- ▶ What if $x = 17$? Interpolate somehow?

Estimation methods

Parametric

- ▶ Distribution is a function with (a few) parameters $\theta = (\theta_1, \theta_2, \dots, \theta_D)$
- ▶ Example: the normal distribution $\mathcal{N}(x|\mu, \sigma^2)$.

Non-parametric

- ▶ Function of *many* parameters.
- ▶ But parameters disappear from estimation methods.
- ▶ Examples: K-nearest neighbours, histogram, Parzen window.

Estimation methods

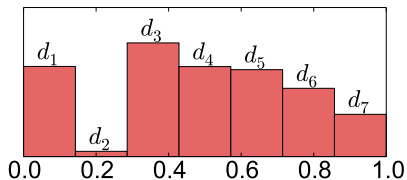
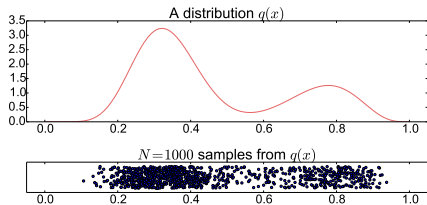
Non-parametric

- ▶ Function of *many* parameters.
- ▶ But parameters disappear from estimation methods.
- ▶ Examples: K-nearest neighbours, histogram, Parzen window.

Histogram as piecewise constant density estimate

Histogram with B bins.

For a given B , the parameters of this piecewise-constant function are the heights d_1, d_2, \dots, d_B of the individual bins. This function is denoted $p(x|\{d_1, d_2, \dots, d_B\})$.



Keep in mind the data x are 1-D. $N = 1000$ samples graph plots them in 1.5-D only because of better visibility.

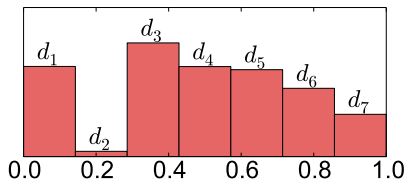
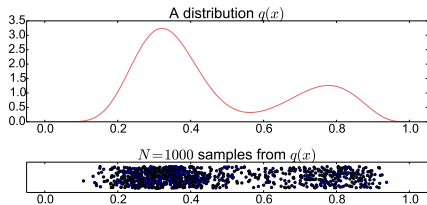
For the given number of bins B , d_1, d_2, \dots, d_B must conform to the constraint that the area under the function must sum up to one,

$$1 = \int_{-\infty}^{\infty} p(x|\{d_1, d_2, \dots, d_B\}) dx = \sum_{i=1}^B \int_{\frac{i-1}{B}}^{\frac{i}{B}} d_i dx = \sum_{i=1}^B d_i \overset{\text{bin width}}{\downarrow} w = \sum_{i=1}^B d_i \frac{1}{B}.$$

Histogram as piecewise constant density estimate

Histogram with B bins.

For a given B , the parameters of this piecewise-constant function are the heights d_1, d_2, \dots, d_B of the individual bins. This function is denoted $p(x|\{d_1, d_2, \dots, d_B\})$.



For the given number of bins B , d_1, d_2, \dots, d_B must conform to the constraint that the area under the function must sum up to one,

$$1 = \int_{-\infty}^{\infty} p(x|\{d_1, d_2, \dots, d_B\}) dx = \sum_{i=1}^B \int_{\frac{i-1}{B}}^{\frac{i}{B}} d_i dx = \sum_{i=1}^B d_i \overset{\text{bin width}}{\downarrow} w = \sum_{i=1}^B \frac{d_i}{B}.$$

Keep in mind the data x are 1-D. $N = 1000$ samples graph plots them in 1.5-D only because of better visibility.

Finding d_j using ML

$$L(\mathcal{T}) = p(\mathcal{T}|\boldsymbol{\theta}) = \prod_{i=1}^N p(x_i|\boldsymbol{\theta}) = \prod_{j=1}^B \overbrace{\left(\prod_{k=1}^{N_j} d_j \right)}^{\text{points in } j\text{-th bin}} = \prod_{j=1}^B d_j^{N_j}.$$

Maximization task:

$$\ell(\mathcal{T}) = \sum_{j=1}^B N_j \log d_j \rightarrow \max, \quad \text{subject to } \frac{1}{B} \sum_{j=1}^B d_j = 1,$$

$$\text{Lagrangian: } \sum_{j=1}^B N_j \log d_j + \lambda \left(\frac{1}{B} \sum_{j=1}^B d_j - 1 \right)$$

$$\frac{N_j}{d_j} + \frac{\lambda}{B} = 0 \Rightarrow \frac{d_j}{N_j} = \text{const.} \Rightarrow d_j = B \frac{N_j}{N}.$$

Finding d_j using ML

$$L(\mathcal{T}) = p(\mathcal{T}|\boldsymbol{\theta}) = \prod_{i=1}^N p(x_i|\boldsymbol{\theta}) = \prod_{j=1}^B \overbrace{\left(\prod_{k=1}^{N_j} d_j \right)}^{\text{points in } j\text{-th bin}} = \prod_{j=1}^B d_j^{N_j}.$$

Maximization task:

$$\ell(\mathcal{T}) = \sum_{j=1}^B N_j \log d_j \rightarrow \max, \quad \text{subject to } \frac{1}{B} \sum_{j=1}^B d_j = 1,$$

$$\text{Lagrangian: } \sum_{j=1}^B N_j \log d_j + \lambda \left(\frac{1}{B} \sum_{j=1}^B d_j - 1 \right)$$

$$\frac{N_j}{d_j} + \frac{\lambda}{B} = 0 \Rightarrow \frac{d_j}{N_j} = \text{const.} \Rightarrow d_j = B \frac{N_j}{N}.$$

Finding d_j using ML

$$L(\mathcal{T}) = p(\mathcal{T}|\boldsymbol{\theta}) = \prod_{i=1}^N p(x_i|\boldsymbol{\theta}) = \prod_{j=1}^B \overbrace{\left(\prod_{k=1}^{N_j} d_j \right)}^{\text{points in } j\text{-th bin}} = \prod_{j=1}^B d_j^{N_j}.$$

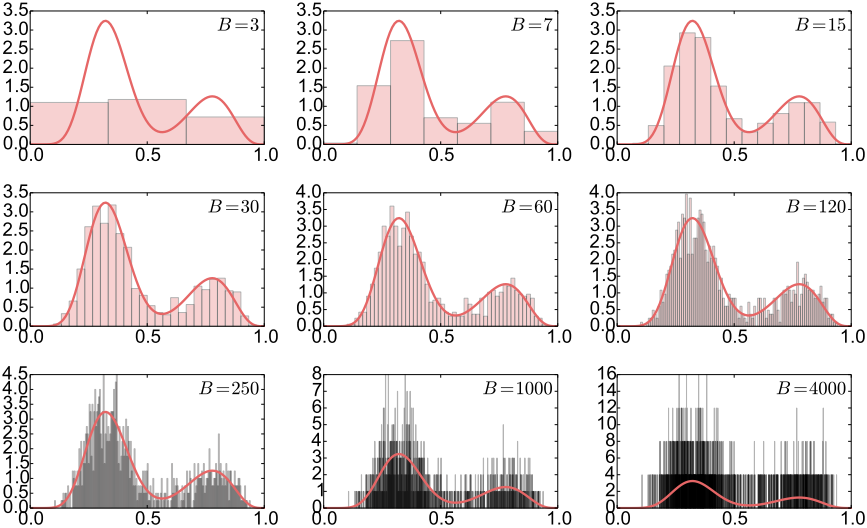
Maximization task:

$$\ell(\mathcal{T}) = \sum_{j=1}^B N_j \log d_j \rightarrow \max, \quad \text{subject to } \frac{1}{B} \sum_{j=1}^B d_j = 1,$$

$$\text{Lagrangian: } \sum_{j=1}^B N_j \log d_j + \lambda \left(\frac{1}{B} \sum_{j=1}^B d_j - 1 \right)$$

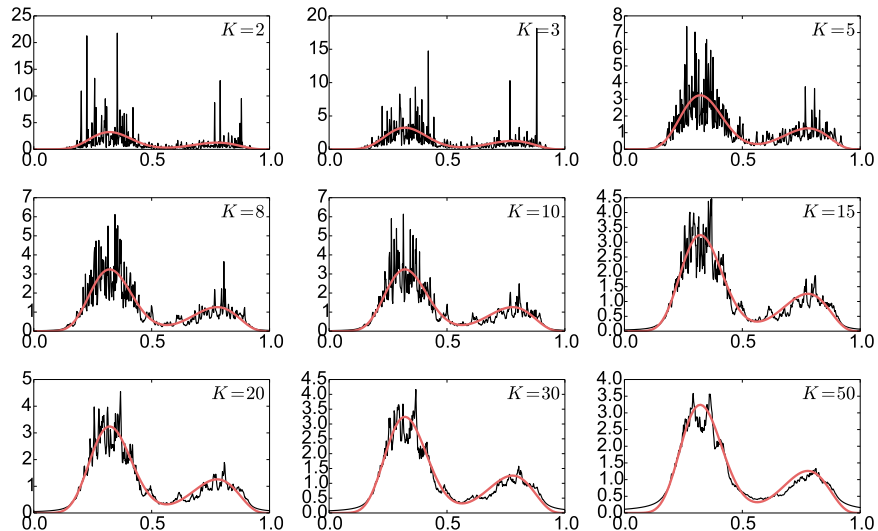
$$\frac{N_j}{d_j} + \frac{\lambda}{B} = 0 \Rightarrow \frac{d_j}{N_j} = \text{const.} \Rightarrow d_j = B \frac{N_j}{N}.$$

Different number of bins



K-Nearest neighbors density estimates

Find K neighbors, the density estimate is then $\rho \sim 1/V$ where V is the volume of a minimum cell containing K NNs.



References I

Further reading: Chapter 13 and 14 of [3]. Books [1] and [2] are classical textbooks in the field of pattern recognition and machine learning. The lecture has been greatly inspired by the 4th and 5th lecture of the Machine Learning and Pattern Recognition course ([B4B33RPZ](#))

[1] Christopher M. Bishop.

Pattern Recognition and Machine Learning.

Springer Science+Business Media, New York, NY, 2006.

[PDF](#) freely downloadable.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork.

Pattern Classification.

John Wiley & Sons, 2nd edition, 2001.

References II

- [3] Stuart Russell and Peter Norvig.
Artificial Intelligence: A Modern Approach.
Prentice Hall, 3rd edition, 2010.
<http://aima.cs.berkeley.edu/>.