Probability estimation

Tomáš Svoboda
thanks to Ondřej Drbohlav, Michal Reinstein, Jiří Matas

Department of Cybernetics, Vision for Robotics and Autonomous Systems,
Center for Machine Perception (CMP)

May 23, 2018

Probabity estimation

In previous two lectures:

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

In practice:

- uknown quantities
- estimate from training data $\mathcal{T} = \{(x_1, s_1), (x_2, s_2), \ldots (x_l, s_l)\}$

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

Problem: Coins classification based on weight

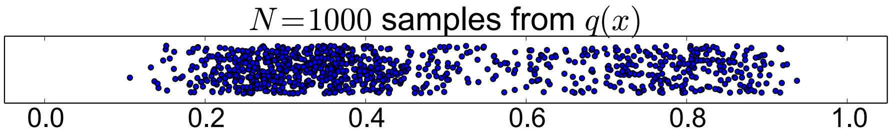
s/x	5 g	10 g	15 g	20 g	25 g	$\sum$
1 CZK	15	10	3	0	0	**28**
2 CZK	7	13	16	6	1	**43**
5 CZK	0	1	2	11	15	**29**
$\sum$	22	24	21	17	16	**100**

- What if $x = 17$? Interpolate somehow?
- Two weighting devices A, B. $x_A = 16$, $x_B = 19$ what to do?

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

Probability (density/distribution) estimation from samples

Training data (for one of the class): $\mathcal{T} = \{x_1, x_2, x_3, \ldots, x_{1000}\}$

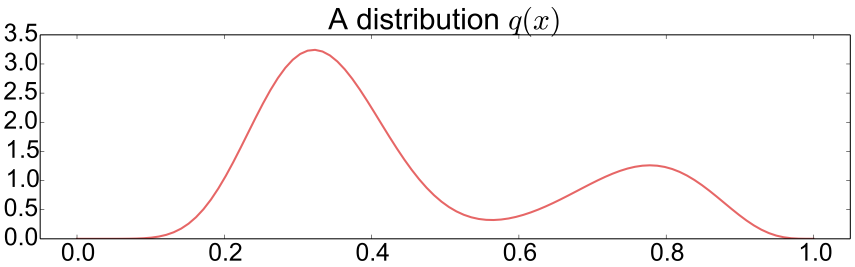


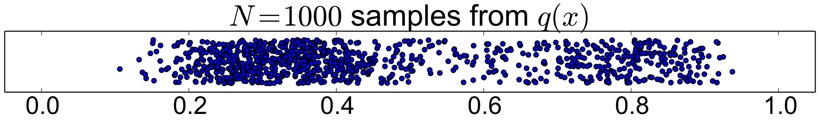
- Range normalized $< 0, 1 >$
- Analysis per class (for each class separately).

Try to draw the density function, guessing from the samples.
The data x inded scalar - the quasi 2D plot is for visualisation, only the x-axis matters. Think about weight feature.
We drop the class index.

About normalization - think about assigning 1 to the max and 0 to min

Probability density/distribution





Estimation methods

Parametric

- Distribution is a function with (a few) parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)$
- 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.

Tossing coin. Likelihood

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

$$P(\mathrm{H,H}|p_H = 0.5) = 0.5^2 = 0.25$$
$$P(\mathrm{H,H}|p_H = 0.2) = 0.2^2 = 0.04$$
$$P(\mathrm{H,H}|p_H = 0.8) = 0.8^2 = 0.64$$

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

Tossed 3×, two heads $\mathcal{T} = \{\mathrm{H,H,T}\}$.
What is p_H?

iid - independent (one toss does not influence the other), identically (the same coin) distributed.
Think about difference between $P(\mathrm{H,H}|p_H)$ vs $P(p_H|\mathrm{H,H})$.
Likelihood $\mathcal{L}$ is not a probability, why?

Tossing coin. Likelihood

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

$$P(\text{H,H}|p_H = 0.5) \quad = 0.5^2 = 0.25$$
$$P(\text{H,H}|p_H = 0.2) \quad = 0.2^2 = 0.04$$
$$P(\text{H,H}|p_H = 0.8) \quad = 0.8^2 = 0.64$$

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

Tossed 3×, two heads $\mathcal{T} = \{\text{H,H,T}\}$.
What is p_H?

iid - independent (one toss does not influence the other), identically (the same coin) distributed.
Think about difference between $P(\text{H,H}|p_H)$ vs $P(p_H|\text{H,H})$.
Likelihood $\mathcal{L}$ is not a probability, why?

Tossing coin. Likelihood

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

$$P(\text{H,H}|p_H = 0.5) = 0.5^2 = 0.25$$
$$P(\text{H,H}|p_H = 0.2) = 0.2^2 = 0.04$$
$$P(\text{H,H}|p_H = 0.8) = 0.8^2 = 0.64$$

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

Tossed $3\times$, two heads $\mathcal{T} = \{\text{H,H,T}\}$.
What is p_H?

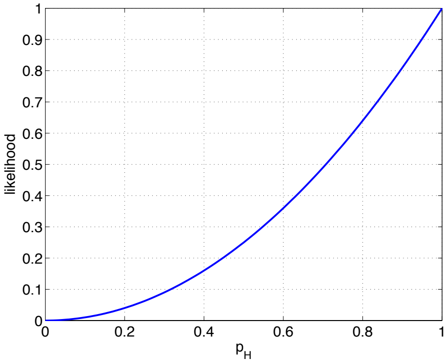
iid - independent (one toss does not influence the other), identically (the same coin) distributed.
Think about difference between $P(\text{H,H}|p_H)$ vs $P(p_H|\text{H,H})$.
Likelihood $\mathcal{L}$ is not a probability, why?

Tossing coin. Likelihood

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

$$P(\text{H,H}|p_H = 0.5) \quad = 0.5^2 = 0.25$$
$$P(\text{H,H}|p_H = 0.2) \quad = 0.2^2 = 0.04$$
$$P(\text{H,H}|p_H = 0.8) \quad = 0.8^2 = 0.64$$

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



Tossed $3\times$, two heads $\mathcal{T} = \{\text{H,H,T}\}$. What is p_H?

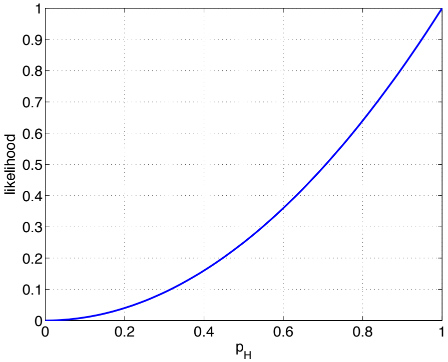
iid - independent (one toss does not influence the other), identically (the same coin) distributed.
Think about difference between $P(\text{H,H}|p_H)$ vs $P(p_H|\text{H,H})$.
Likelihood $\mathcal{L}$ is not a probability, why?

Tossing coin. Likelihood

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

$$
\begin{aligned}
P(\text{H,H}|p_H = 0.5) &= 0.5^2 = 0.25 \\
P(\text{H,H}|p_H = 0.2) &= 0.2^2 = 0.04 \\
P(\text{H,H}|p_H = 0.8) &= 0.8^2 = 0.64
\end{aligned}
$$

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



Tossed 3×, two heads $\mathcal{T} = \{\text{H,H,T}\}$. What is p_H?

iid - independent (one toss does not influence the other), identically (the same coin) distributed.

Think about difference between $P(\text{H,H}|p_H)$ vs $P(p_H|\text{H,H})$.

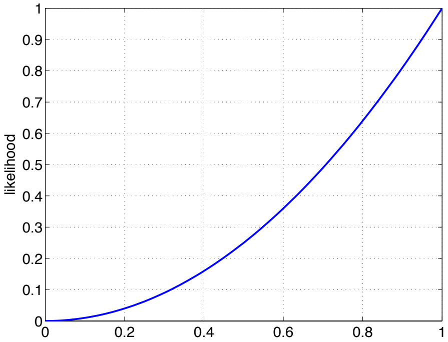
Likelihood $\mathcal{L}$ is not a probability, why?

Tossing coin. Likelihood

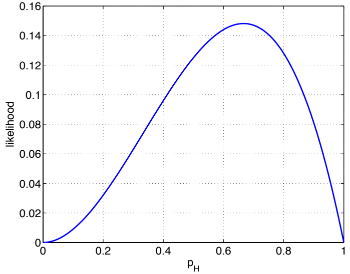
Tossed 2×, two heads $\mathcal{T} = \{\text{H,H}\}$.
We assume iid.

$$
\begin{aligned}
P(\text{H,H}|p_H = 0.5) &= 0.5^2 = 0.25 \\
P(\text{H,H}|p_H = 0.2) &= 0.2^2 = 0.04 \\
P(\text{H,H}|p_H = 0.8) &= 0.8^2 = 0.64
\end{aligned}
$$

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



Tossed 3×, two heads $\mathcal{T} = \{\text{H,H,T}\}$.
What is p_H?



iid - independent (one toss does not influence the other), identically (the same coin) distributed.
Think about difference between $P(\text{H,H}|p_H)$ vs $P(p_H|\text{H,H})$.
Likelihood $\mathcal{L}$ is not a probability, why?

Tossing coin, Maximum likelihood estimate

$x_n = 1$ 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 $n = 1$.

Maximum Likelihood (ML)

Observations $\mathcal{T} = \{x_1, x_2, x_3, \ldots, x_N\}$; known parametric form of the likelihood function $\mathcal{L}(\boldsymbol{\theta}) = p(\mathcal{T}|\boldsymbol{\theta})$.

Maximum likelihood estimate:

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{T}|\boldsymbol{\theta})$$

We assume independent and identically distributed (i.i.d) samples x in $\mathcal{T}$.

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})$$

We can do log-likelihood (logarithm is an increasing function).

$p(\mathcal{T}|\boldsymbol{\theta})$ likelihood that the data $\mathcal{T}$ were generated by the density/distribution function with parameters $\boldsymbol{\theta}$. If parameters are correct they will do larger probabilites (hence the max) compared to the wrong ones

Independent - we can use the product of individual probabilies

Identically - from the same distribution

Example: Normal distribution

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

$$p(\{x_1, x_2, \ldots, x_N\}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})$$

$$p(\mathcal{T}|\mu,\sigma) = \frac{1}{\sigma^N\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right]$$

We are looking for an extremum of $p(\mathcal{T}|\mu,\sigma)$

Derivation on the blackboard, or by yourself. You can also logarithm the whole thing.

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2$$

Why the Normal distribution

Central Limit Theorem

$$X = X_A + X_B + X_C$$

$X_{A,B,C}$ random variables with uniform distributions

see the `clt.m` demo

Does ML solve it all?

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

Does ML solve it all?

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

Does ML solve it all?

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

Tossing coin, using priors

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

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

(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

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

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

(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

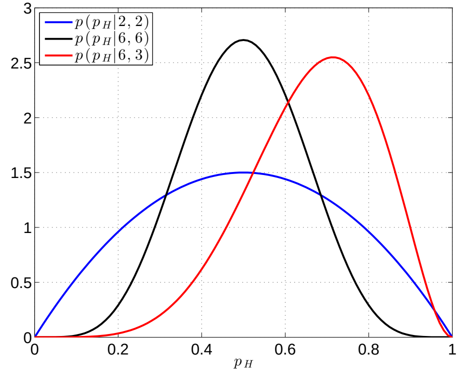
Tossing coin, using priors

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

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

(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

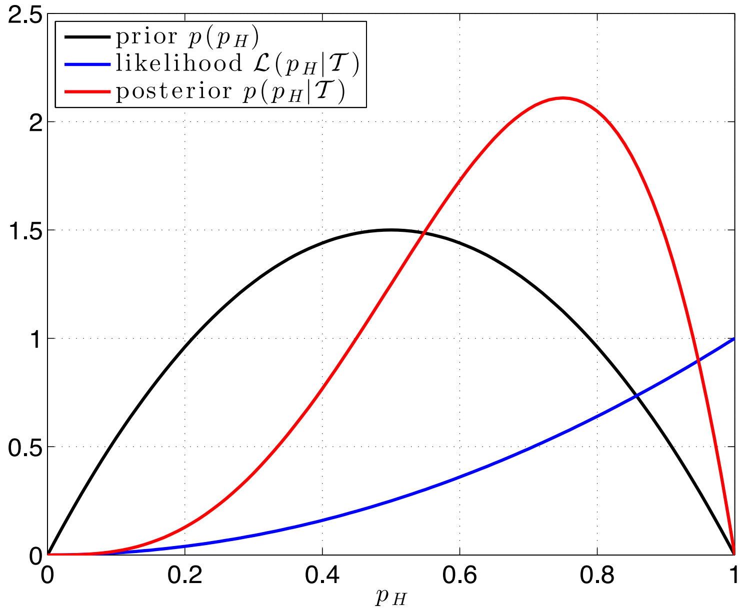
$$\frac{\partial p(p_H|h, N)}{\partial p_H} = 0$$

yields

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

Hyperparamaters a, b as regularization

Maximum aposteriori estimate



See the `map.m` demo.

Estimation methods

Parametric

- Distribution is a function with (a few) parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)$
- 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

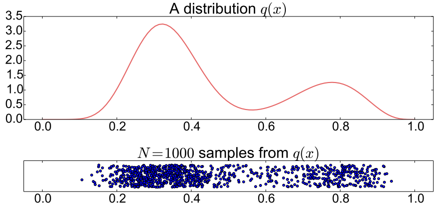
Non-parametric

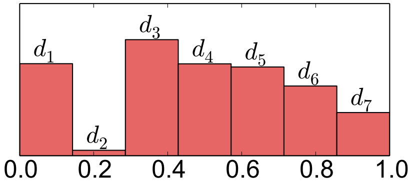
- Function of *many* parameters.
- But parameters disappear from estimation methods.
- Examples: K-nearest neighbours, histogram, Parzen window.

Histogram as piecewise constant density estimate

Histogram with B bins.

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





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

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

Finding d_i 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:

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

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

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

Finding d_i 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:

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

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

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

Finding d_i using ML

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

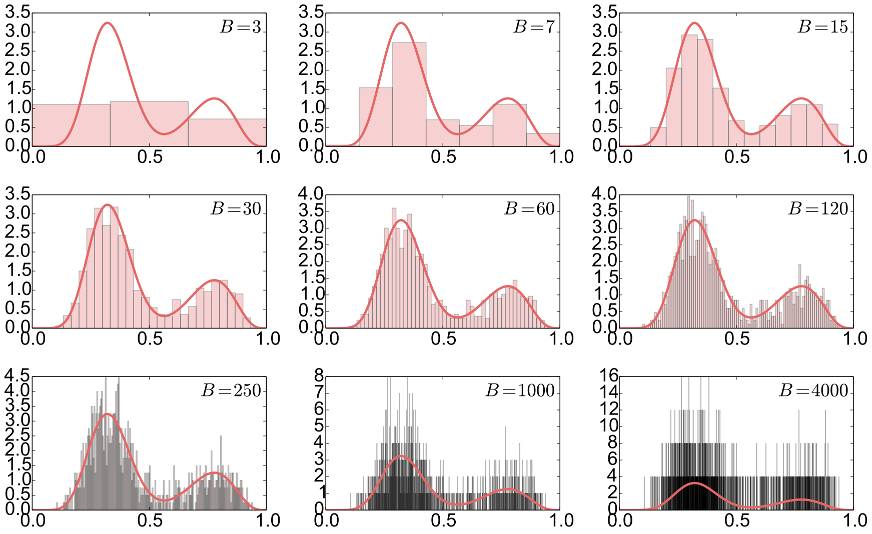
Maximization task:

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

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

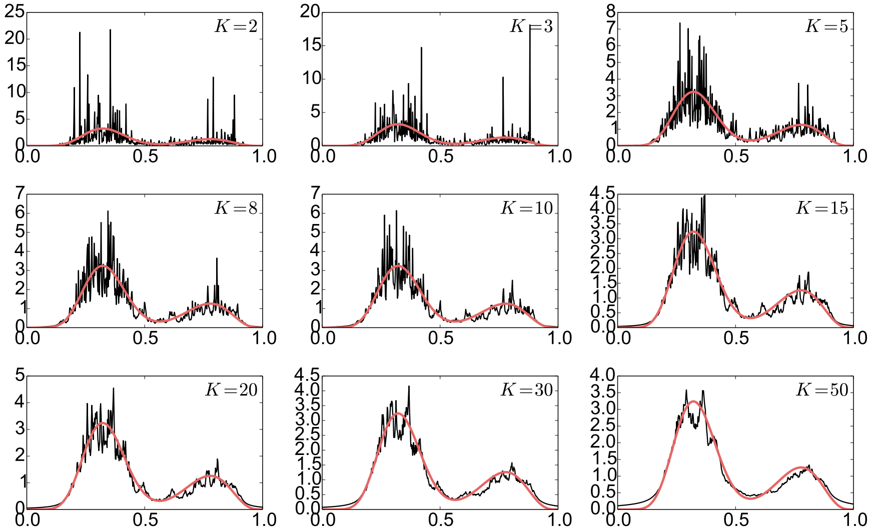
$$\frac{N_j}{d_j} + \frac{\lambda}{B} = 0 \Rightarrow \frac{d_j}{N_j} = \text{const.} \Rightarrow \boxed{d_j = B\frac{N_j}{N}}\,.$$

Different number of bins



K-Nearest neighbors density estimates

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



Maximum likelihood estimation

(Back to the coin example) 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:

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

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

after some derivation, ..., weighted average

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

Maximum likelihood estimation

(Back to the coin example) 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:

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

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

after some derivation, . . . , weighted average

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

Maximum likelihood estimation

(Back to the coin example) 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:

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

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

after some derivation, …, weighted average

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

Maximum likelihood estimation

(Back to the coin example) 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:

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

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

after some derivation, ..., weighted average

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

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+Bussiness Media, New York, NY, 2006.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork.
Pattern Classification.
John Wiley & Sons, 2nd edition, 2001.

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