Bayesian Models in Machine Learning

Lukáš Burget

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Escuela de Ciencias Informáticas 2017

Buenos Aires, July 24-29 2017

Frequentist vs. Bayesian

- Frequentist point of view:
  - Probability is the frequency of an event occurring in a large (infinite) number of trials
  - E.g. When flipping a coin many times, what is the proportion of heads?
- Bayesian
  - Inferring probabilities for events that have never occurred or believes which are not directly observed
  - Prior believes are specified in terms of prior probabilities
  - Taking into account uncertainty (posterior distribution) of the estimated parameters or hidden variables in our probabilistic model.

Simple classification problem – I.

- Simple example of learning a probabilistic model for maximum a-posteriori classification
  - to introduce classification as a basic problem from machine learning field
  - to understand frequentist’s view of “probability” and to show its limitations as compared to the Bayesian approaches
  - to refresh basics from probability theory
- The task is to classify an object (*grenade* or *apple*) given an observation (discrete weight category)
  - It is heavy. Is it grenade or apple?
- Lets have 150 observations as training data
  - Table of observation counts for each class and weight category

	$lightest$	$lighter$	$light$	$middle$	$heavy$	$heavier$	$heaviest$	
grenade	1	6	12	15	12	2	2	50
apple	4	22	50	14	6	3	1	100
	0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Simple classification problem – II.

- Lets estimate joint probabilities $P(class, observation)$
  - normalizing the counts by the total count gives Maximum likelihood (ML) estimates (see later): $P(grenade, heavy) = \frac{12}{150}$
  - We need many observations to obtain robust estimates this way.
  - How certain can we be about correctness of these estimates?

- Maximum a-posteriori classification rule:
  - given an observation select the most likely class
  - i.e. select class with highest posterior probability $P(class|observation)$
  - ML estimate: $P(grenade|heavy) = \frac{12}{12+6}$

$\frac{1}{150}$	$\frac{6}{150}$	$\frac{12}{150}$	$\frac{15}{150}$	$\frac{12}{150}$	$\frac{2}{150}$	$\frac{2}{150}$	$\frac{50}{150}$
$\frac{4}{150}$	$\frac{22}{150}$	$\frac{50}{150}$	$\frac{14}{150}$	$\frac{6}{150}$	$\frac{3}{150}$	$\frac{1}{150}$	$\frac{100}{150}$
$lightest$	$lighter$	$light$	$middle$	$heavy$	$heavier$	$heaviest$	
0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Basic rules of probability theory – I.

Sum rule:

$$P(x) = \sum_y P(x, y)$$

Product rule:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

Bayes rule:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Basic rules of probability theory – II.

- Sum rule:

$$P(heavy) = P(grenade, heavy) + P(apple, heavy) = \frac{12}{150} + \frac{6}{150} = \frac{18}{150}$$

$$P(grenade) = \sum_x P(grenade, x) = \frac{50}{150}$$

- Product rule:

$$P(grenade, heavy) = P(grenade|heavy)P(heavy) = \frac{12}{18}\frac{18}{150} = \frac{12}{150}$$

$$P(grenade, heavy) = P(heavy|grenade)P(grenade) = \frac{12}{50}\frac{50}{150} = \frac{12}{150}$$

	$\frac{1}{150}$	$\frac{6}{150}$	$\frac{12}{150}$	$\frac{15}{150}$	$\frac{12}{150}$	$\frac{2}{150}$	$\frac{2}{150}$	$\frac{50}{150}$
	$\frac{4}{150}$	$\frac{22}{150}$	$\frac{50}{150}$	$\frac{14}{150}$	$\frac{6}{150}$	$\frac{3}{150}$	$\frac{1}{150}$	$\frac{100}{150}$
	lightest	*lighter*	*light*	*middle*	*heavy*	*heavier*	*heaviest*	
	0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

Basic rules of probability theory – III.

- Bayes rule:

Posterior probability — Likelihood — Prior probability

$$P(grenade|heavy) = \frac{P(heavy|grenade)P(grenade)}{P(heavy)}$$

Evidence

- The evidence can be evaluated using the sum and product rules in terms of likelihoods and priors:

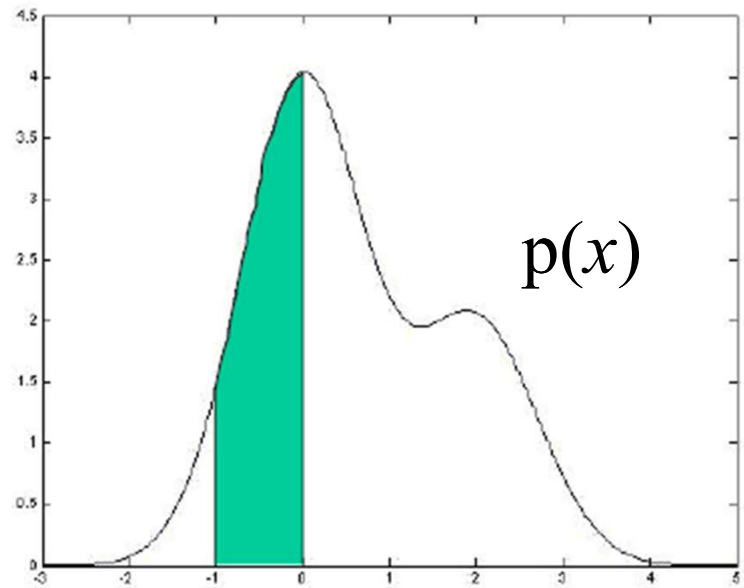
$$P(heavy) = P(heavy|grenade)P(grenade) + P(heavy|apple)P(apple)$$

- Bayes rule for calculating the class posterior may not seem very useful now, but it will be useful in case continuous valued observations.

Continuous random variables

- $P(x)$ –probability
- $p(x)$ –probability density function

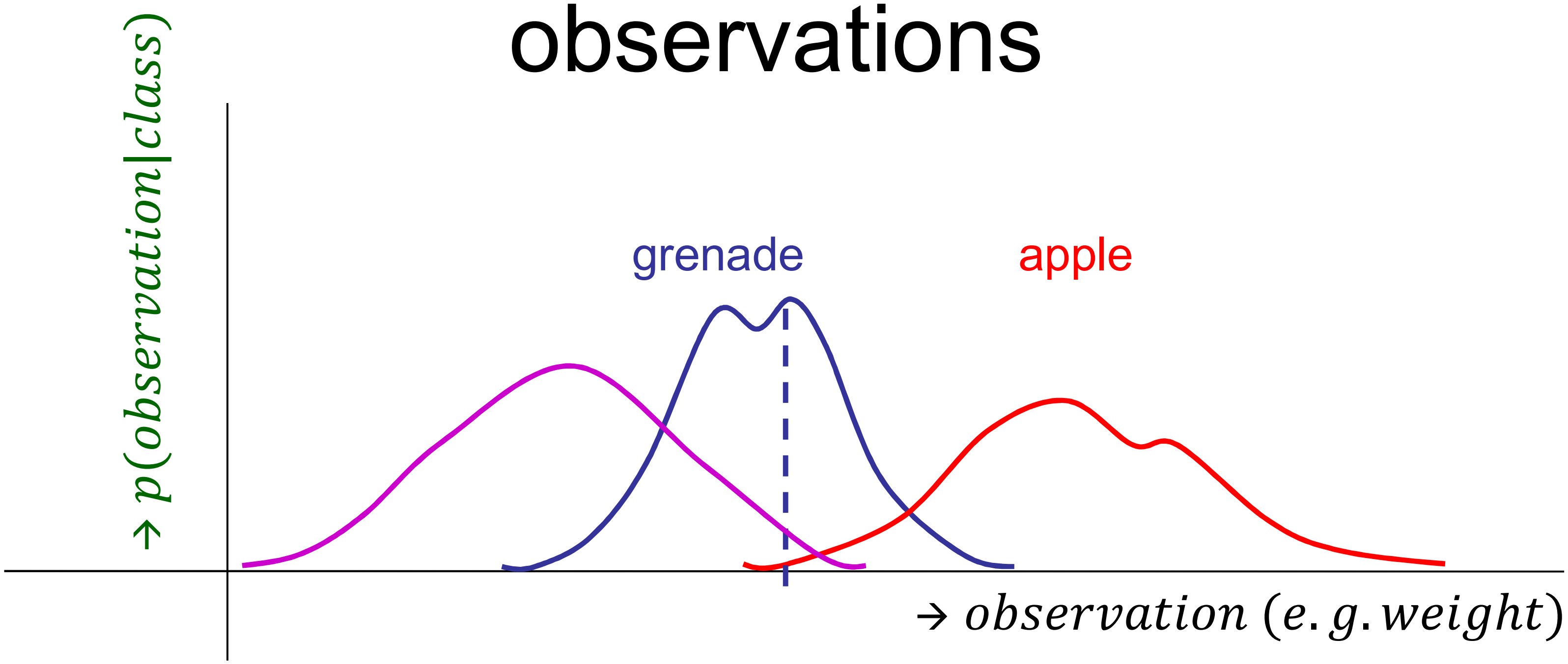
$$P(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$



Sum rule:

$$p(x) = \int p(x,y)\,dy$$

Classification with continuous observations



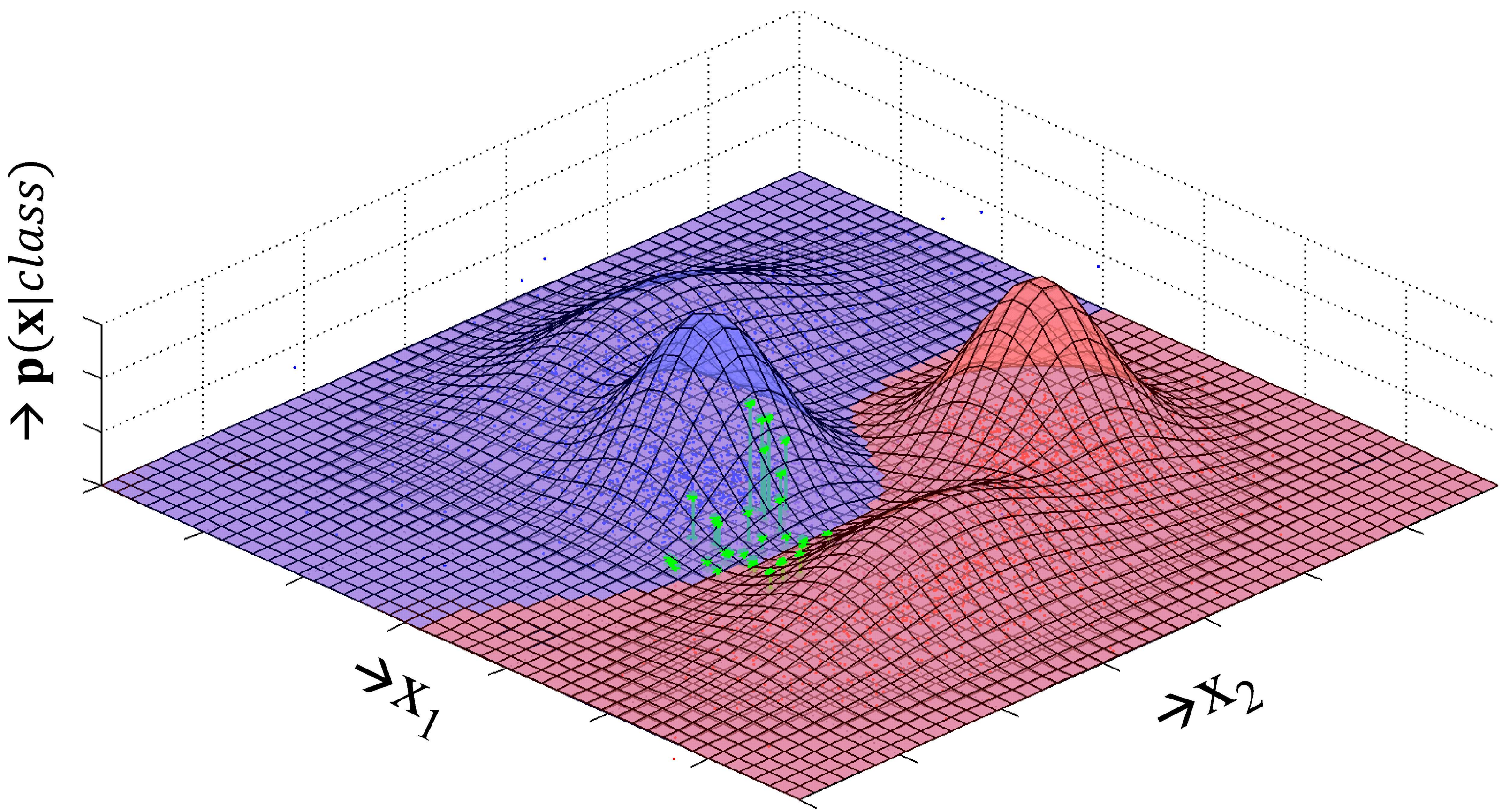
- Maximum a-posteriori classification rule says: select the more likely class

$$P(class|observation) = \frac{p(observation|class)P(class)}{p(observation)}$$

$$P(observation) = \sum_{class} p(observation|class)P(class)$$

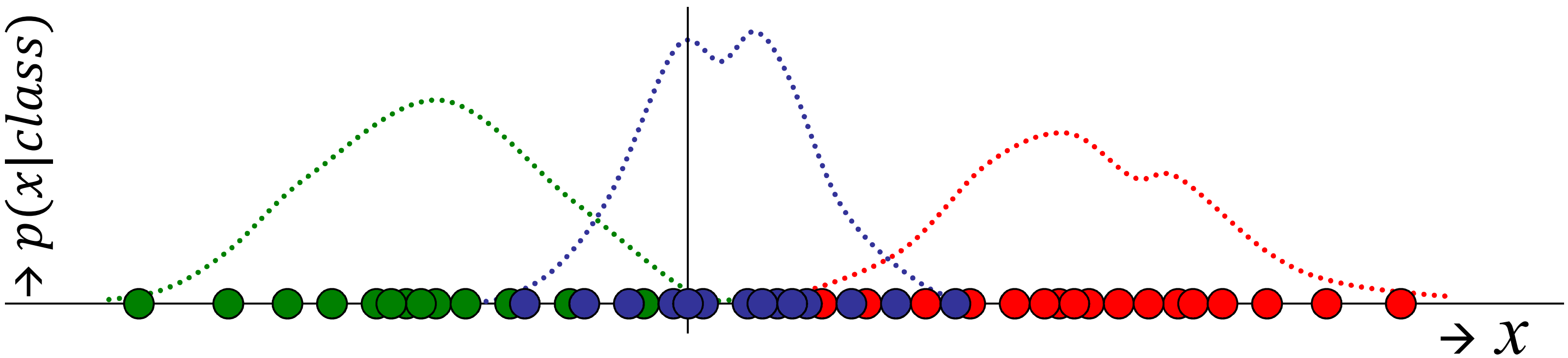
Multivariate observations

From now, univariate observations will be denoted as x and multivariate as $\mathbf{x} = [x_1, x_2, \dots x_D] = [weight, diameter, \dots]$



Estimation of parameters

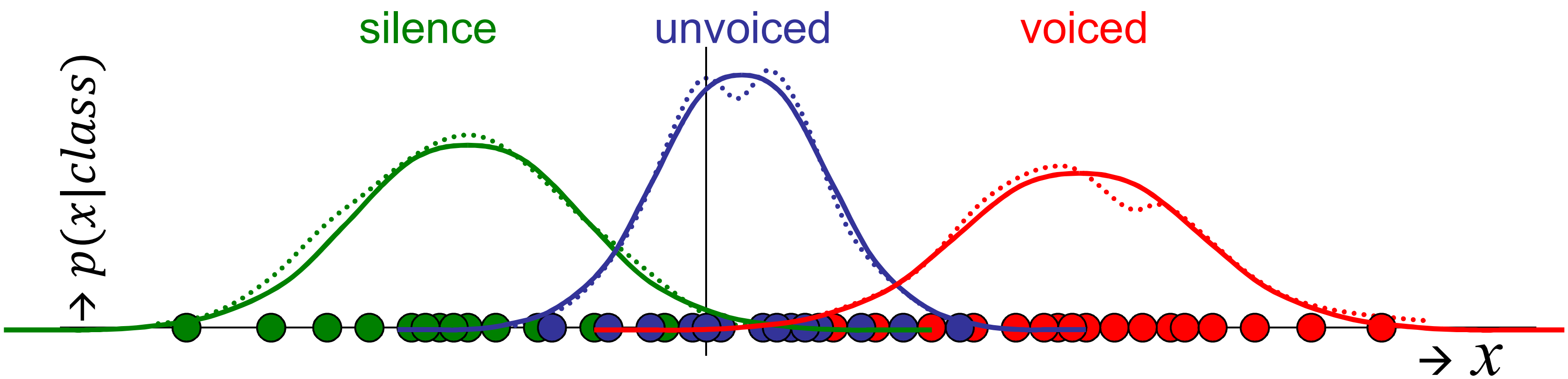
- Usually we do not know the true distributions $p(x|class)$



Estimation of parameters

… we only see some training examples.

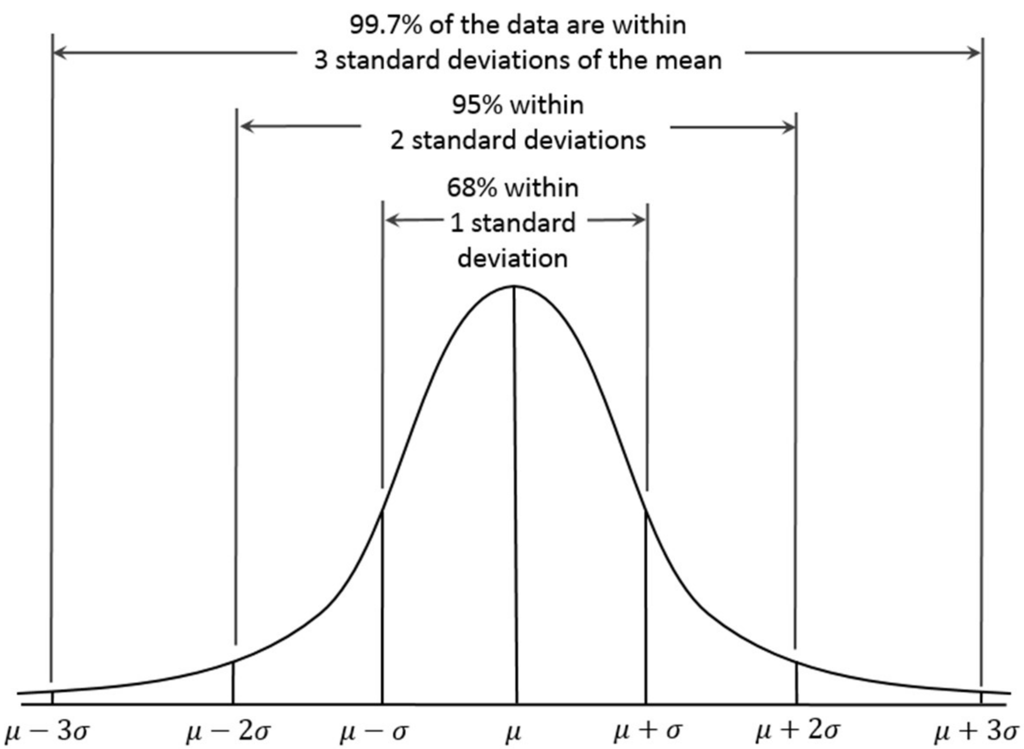
- Let's decide for some parametric model for $p(x|class)$ (e.g. Gaussian distribution) and estimate its parameters from the data.



- Here, we are using the **frequentist approach**: Estimated distributions tell us that observation x will be more likely as we see more similar observations in the training data.
- From now, lets forget about classes. We will concentrate just on estimating probability density functions (e.g. one for each class).

Gaussian distribution (univariate)

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



ML estimates of parameters

$$\mu = \frac{1}{N}\sum_n x_n$$

$$\sigma^2 = \frac{1}{N}\sum_n (x_n - \mu)^2$$

Why Gaussian distribution?

- Simple and easy to deal with
  - Just a quadratic function in log domain

$$\log \mathcal{N}(x; \mu, \sigma^2) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}(x-\mu)^2 = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} + K$$

  - Likelihood of observed sequence $\mathbf{x} = [x_1, x_2, x_3, \ldots x_N]$ is

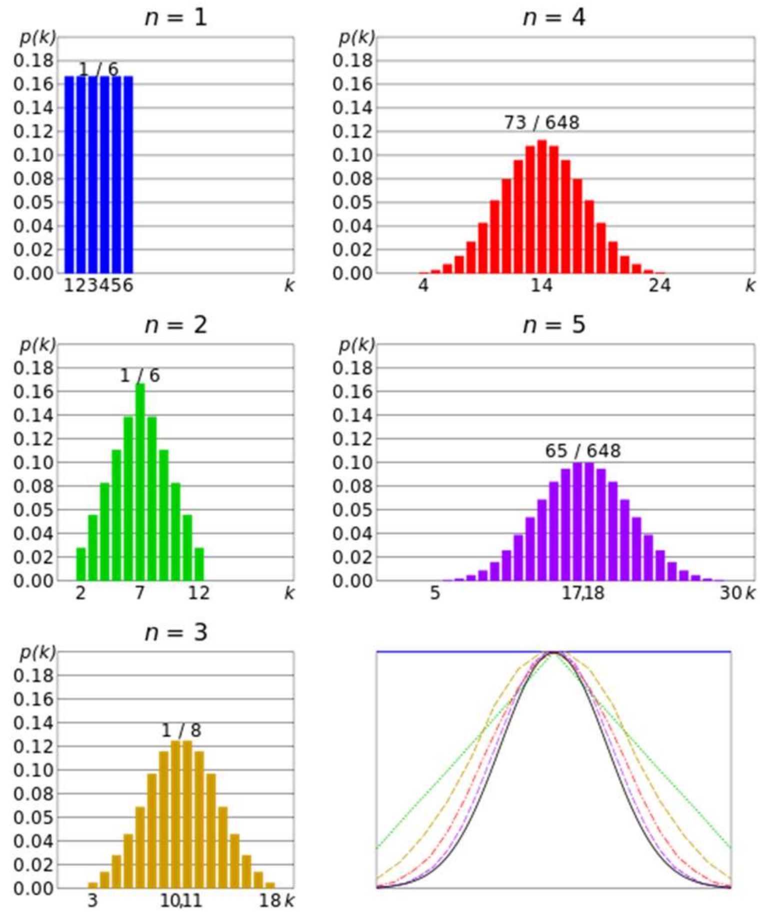
$$p(\mathbf{x}|\mu, \sigma^2) = \prod_n \mathcal{N}(x_n; \mu, \sigma^2) = \exp\left\{\sum_n \log \mathcal{N}(x_n; \mu, \sigma^2)\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\sum_n x_n^2 + \frac{\mu}{\sigma^2}\sum_n x_n - N\frac{\mu^2}{2\sigma^2} + NK\right\}$$

Sufficient statistics
(second, first and zero order)

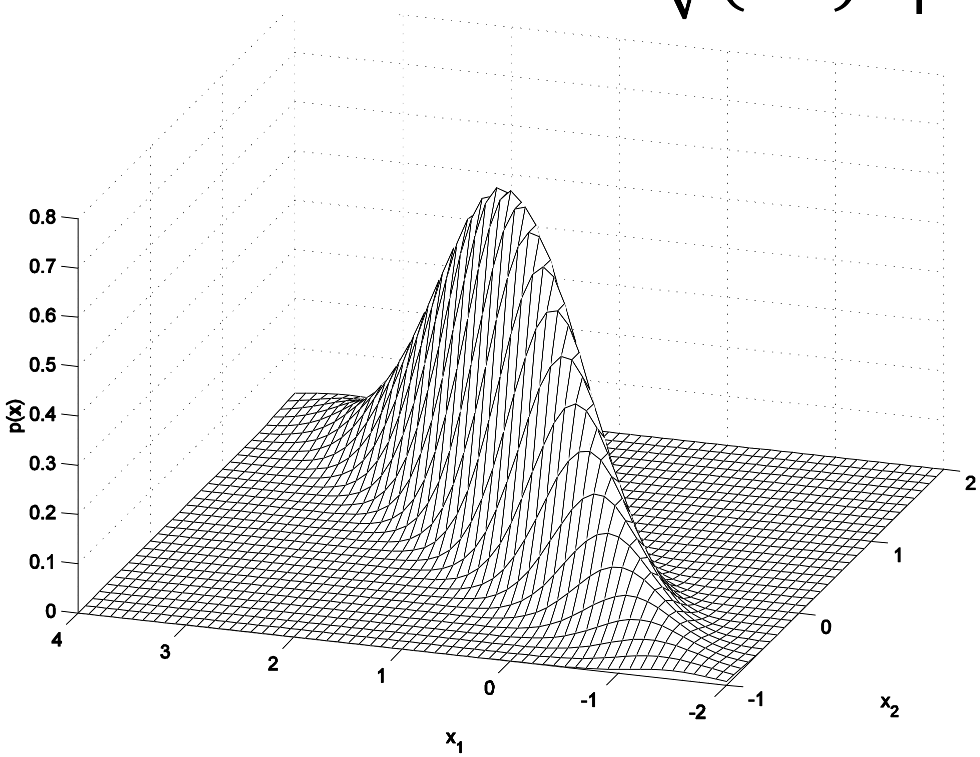
Why Gaussian distribution?

- Naturally occurring
- Central limit theorem: Summing values of many independently generated random variables gives Gaussian distributed observations
- Examples:
  - Summing outcome of N dices
  - Galton's board
    https://www.youtube.com/watch?v=03tx4v0i7MA



Gaussian distribution (multivariate)

$$p(x_1, \ldots, x_D) =$$
$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$



ML estimates of parameters

$$\boldsymbol{\mu} = \frac{1}{N} \sum_n \mathbf{x}_n$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_n (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T$$

Maximum likelihood estimation of parameters

- Lets choose a parametric distribution $p(\mathbf{x}|\boldsymbol{\eta})$ with parameters $\boldsymbol{\eta}$
  - Gaussian distribution with parameters μ, σ^2
- … and lets have some observed training data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$, which we assume to be i.i.d. generated from this distribution.
- We might obtain maximum likelihood estimates of the parameters $\widehat{\boldsymbol{\eta}}^{ML}$ by maximizing the likelihood of the observed data

$$\widehat{\boldsymbol{\eta}}^{ML} = \arg\max_{\boldsymbol{\eta}} p(\mathbf{X}|\boldsymbol{\eta}) = \arg\max_{\boldsymbol{\eta}} \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\eta})$$

- Later, we will see that, under some assumptions, this estimates gives us the most likely parameters.

ML estimate for Gaussian

$$\arg\max_{\mu,\sigma^2} p(\mathbf{x}|\mu,\sigma^2) = \arg\max_{\mu,\sigma^2} \log p(\mathbf{x}|\mu,\sigma^2) = \arg\max_{\mu,\sigma^2} \sum_n \log \mathcal{N}(x_n;\mu,\sigma^2)$$

$$= \arg\max_{\mu,\sigma^2} \left(-\frac{1}{2\sigma^2} \sum_n x_n^2 + \frac{\mu}{\sigma^2} \sum_n x_n - N\frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi)}{2} \right)$$

$$\frac{\partial}{\partial\mu} \log p(\mathbf{x}|\mu,\sigma^2) = \frac{\partial}{\partial\mu} \left(-\frac{1}{2\sigma^2} \sum_n x_n^2 + \frac{\mu}{\sigma^2} \sum_n x_n - N\frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi)}{2} \right)$$

$$= \frac{1}{\sigma^2} \left(\sum_n x_n - N\mu \right) = 0 \quad \Rightarrow \quad \hat{\mu}^{ML} = \frac{1}{N} \sum\nolimits_n x_n$$

and similarly: $\widehat{\sigma^2}^{ML} = \frac{1}{N} \sum\nolimits_n (x_n - \mu)^2$

Categorical distribution

4	22	50	14	6	3	1	100
$lightest$	$lighter$	$light$	$middle$	$heavy$	$heavier$	$heaviest$	
0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.7	[kg]

$$p(x|\boldsymbol{\pi}) = \mathrm{Cat}(x|\boldsymbol{\pi}) = \pi_x$$

- Also referred to as Discrete distribution
- Special binary case is Bernoulli distribution
- $x \in \{lightest, lighter, light, middle, heavy, heavier, heaviest\}$
  or x can be simply the index of a category $\mathbf{x} \in \{1,2,\dots,C\}$
- $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_C]$ - probabilities of the categories are the parameters
- Likelihood of an observed training set $\mathbf{x} = [x_1, x_2, \dots, x_N]$

$$P(\mathbf{x}|\boldsymbol{\pi}) = \prod_n \mathrm{Cat}(\mathbf{x}_n|\boldsymbol{\pi}) = \prod_n \pi_{x_n} = \prod_c \pi_c^{m_c}$$

where m_c is number of observations from category c.

  – (e.g. the numbers from the table)

ML estimate for Categorical

$$\arg\max_{\boldsymbol{\pi}} p(\mathbf{x}|\boldsymbol{\pi}) = \arg\max_{\boldsymbol{\pi}} \log p(\mathbf{x}|\boldsymbol{\pi}) = \arg\max_{\boldsymbol{\pi}} \log \prod_{n=1}^{N} \mathrm{Cat}(x_n|\boldsymbol{\pi})$$

$$= \arg\max_{\boldsymbol{\pi}} \log \prod_{c} \pi_c^{m_c} = \arg\max_{\boldsymbol{\pi}} \sum_{c} m_c \log \pi_c$$

We need to use Lagrange multiplier λ to enforce the constraint $\sum_k \pi_k = 1$

$$\frac{\partial}{\partial \pi_c} \log p(\mathbf{x}|\boldsymbol{\pi}) = \frac{\partial}{\partial \pi_c}\left(\sum_k m_k \log \pi_k - \lambda\left(\sum_k \pi_k - 1\right)\right) = m_c - \lambda = 0$$

$$\Rightarrow \pi_c = \frac{m_c}{\lambda} = \frac{m_c}{N}$$