Probability and Estimation

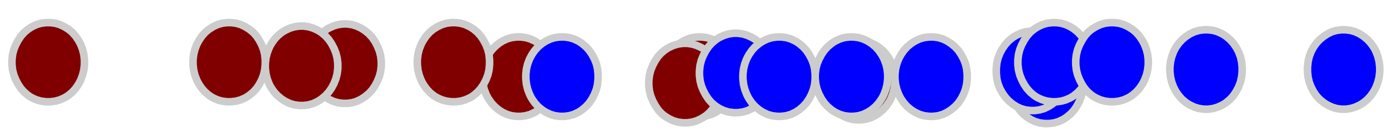
StatML 6.2.2014

Aasa Feragen

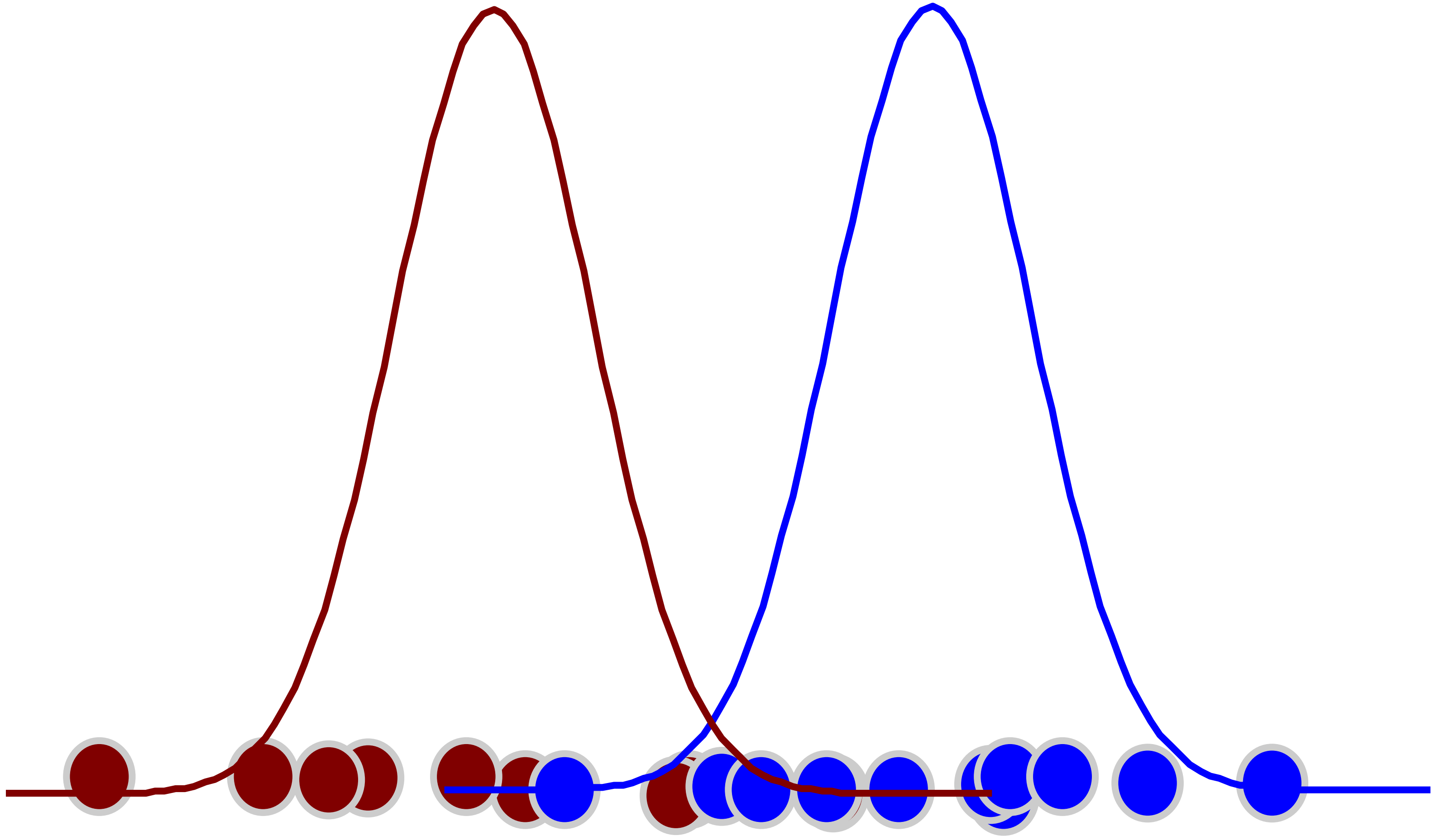
After today's lecture you should

- Know the theoretical background for estimation of distributions
- Know the principles of Bayesian estimation
- Know standard techniques for parametric and non-parametric estimation of probability distributions
  - Maximum likelihood and maximum a posteriori estimation
  - Examples of non-parametric methods (more to come later in the course)
  - Conjugate priors
- Be able to use the above parametric techniques for estimation of Gaussian distributions in real problems
- You will meet these topics in Assignments 1 and 2!

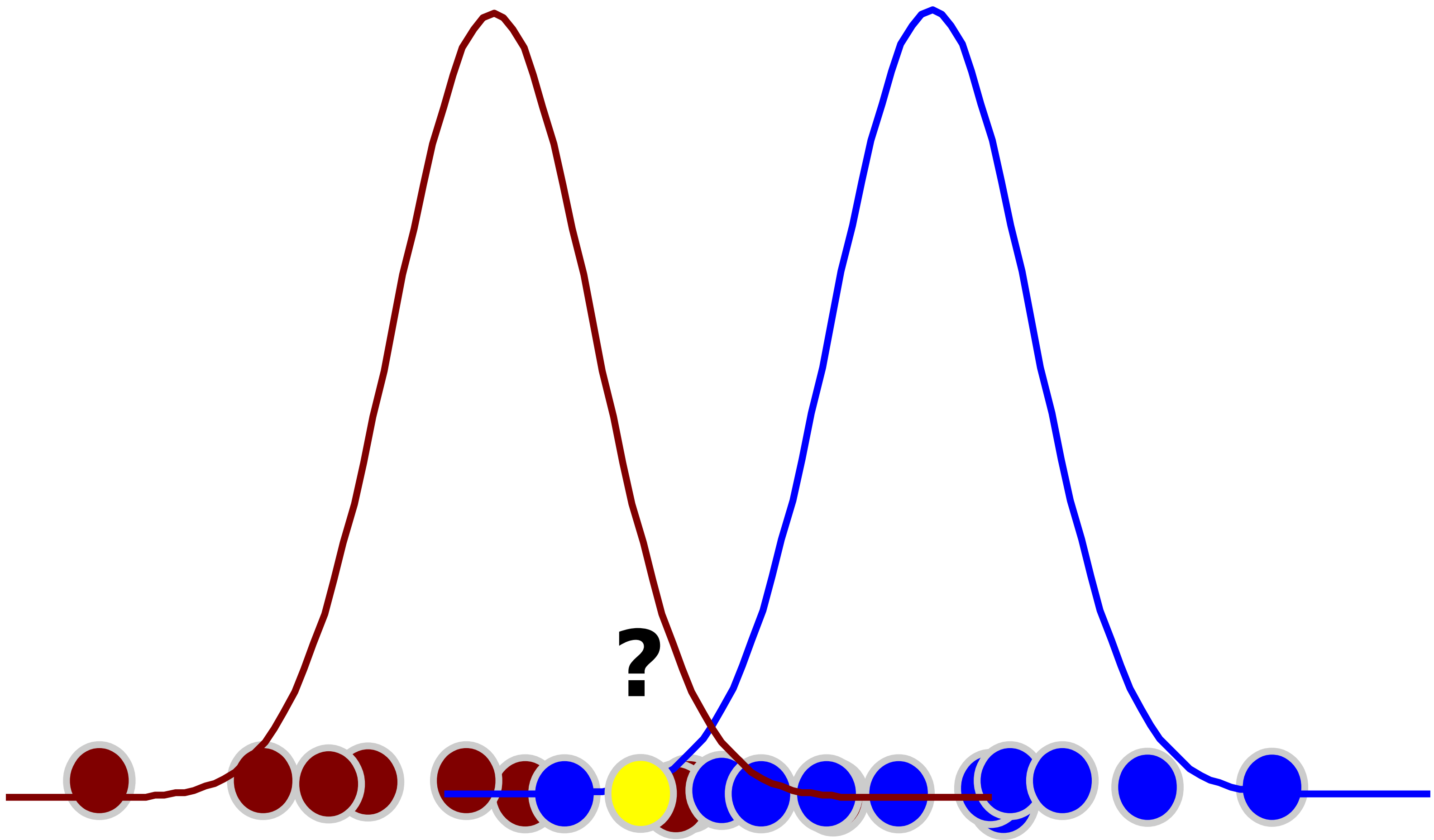
Recall: Probability distributions important for probabilistic ML...



Recall: Probability distributions important for probabilistic ML...



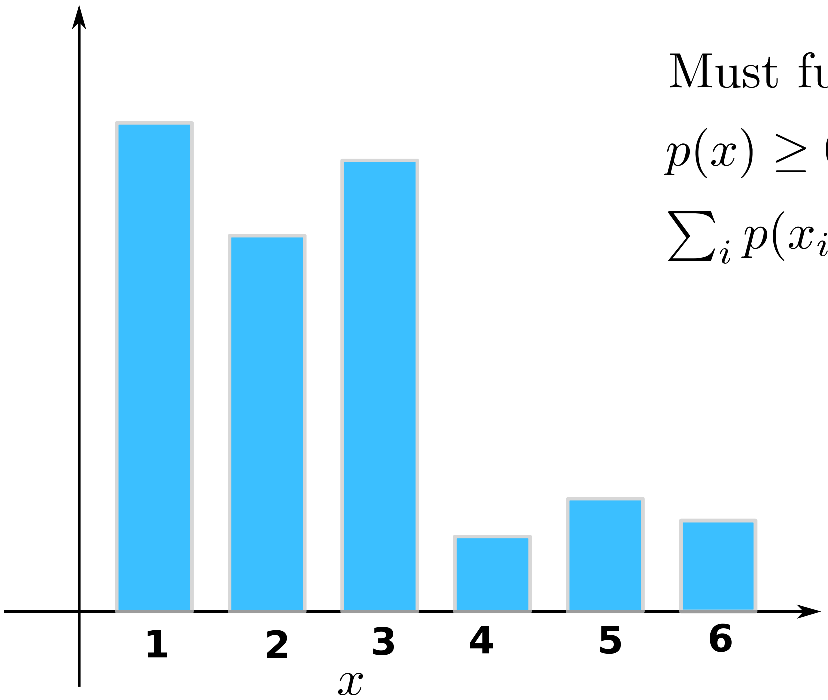
Recall: Probability distributions important for probabilistic ML...



Recall from last time!

Discrete random variables:

$p(x) = p(X = x)$ is called a *probability mass function*



Must fulfill

$p(x) \geq 0$ for all x

$\sum_i p(x_i) = 1$

Recall from last time: Continuous random variables

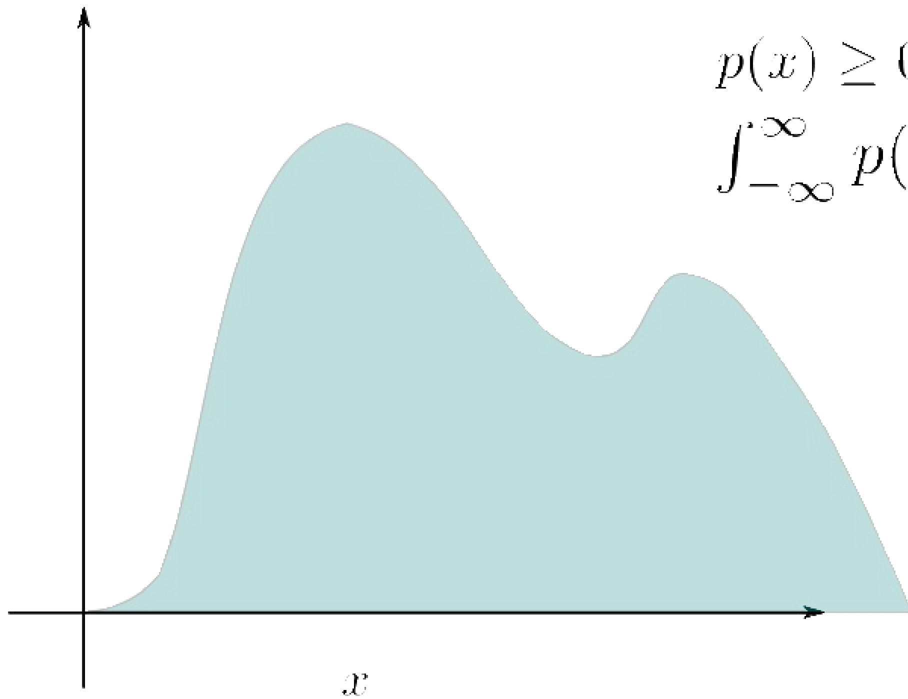
$x \in \mathbb{R}$ real random variable

$p \colon \mathbb{R} \to \mathbb{R}$

Must fulfill

$p(x) \geq 0$ for all x

$\int_{-\infty}^{\infty} p(x)dx = 1$



Recall from last time: Continuous random variables

$x \in \mathbb{R}$ real random variable

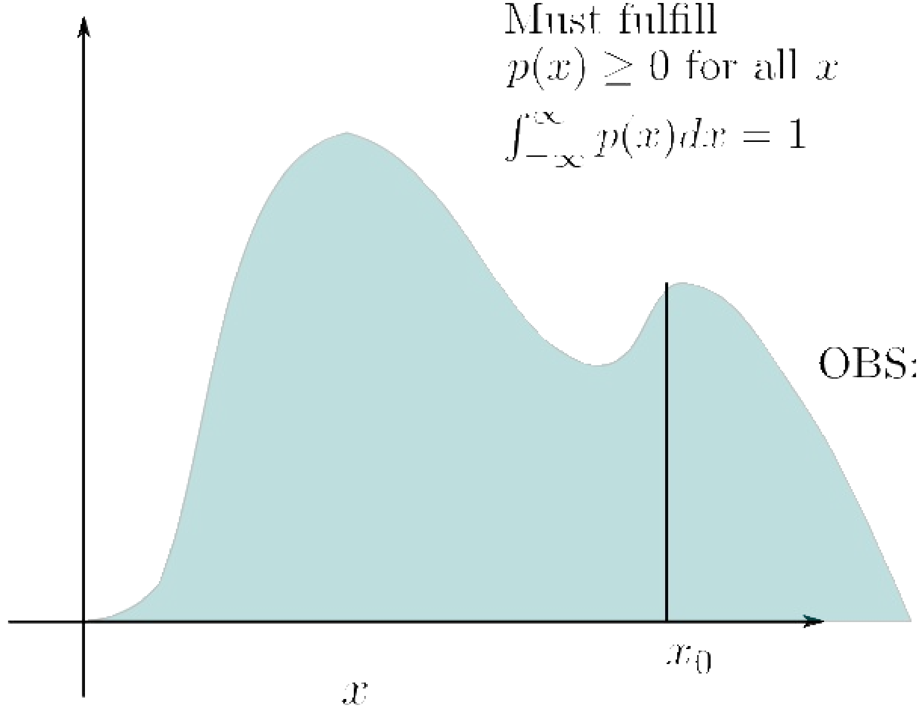
$p \colon \mathbb{R} \to \mathbb{R}$

$p(x)$ is the *probability density function* of X

Must fulfill

$p(x) \geq 0$ for all x

$\int_{-\infty}^{\infty} p(x)dx = 1$



OBS: $p(X = x_0) = 0$

Recall from last time: Continuous random variables

$x \in \mathbb{R}$ real random variable

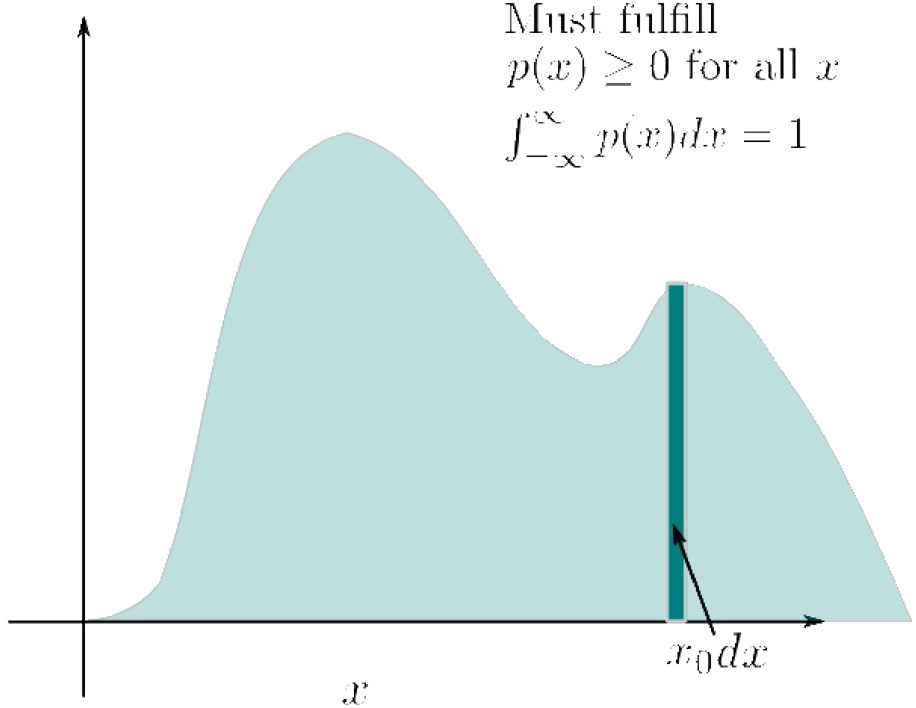
$p: \mathbb{R} \to \mathbb{R}$

$p(x)$ is the *probability density function* of X

Must fulfill

$p(x) \geq 0$ for all x

$\int_{-\infty}^{\infty} p(x)dx = 1$



Recall from last time: Continuous random variables

$x \in \mathbb{R}$ real random variable

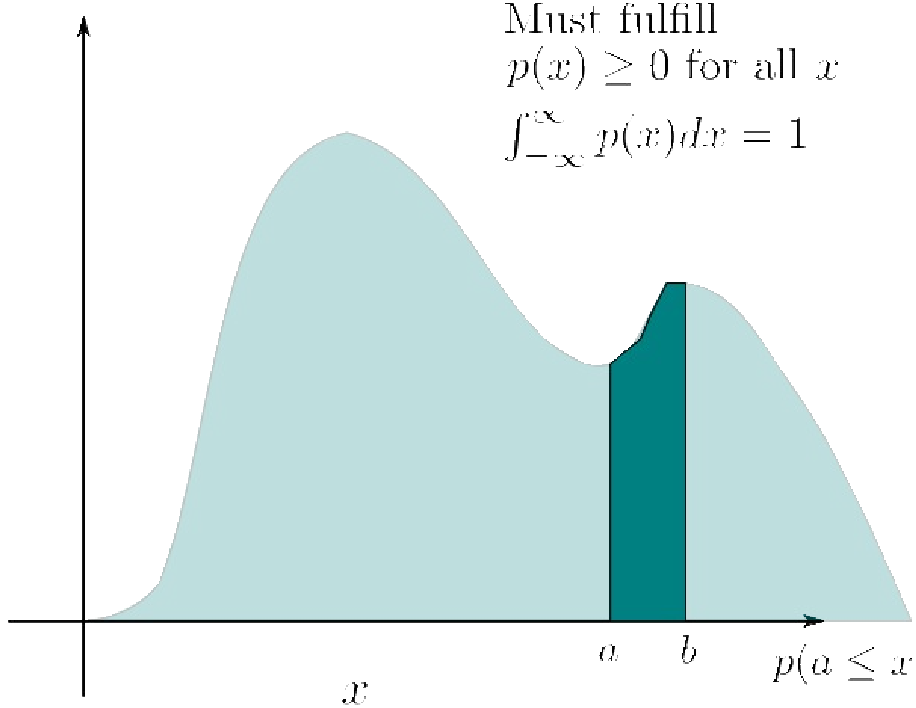
$p \colon \mathbb{R} \to \mathbb{R}$

$p(x)$ is the *probability density function* of X

Must fulfill

$p(x) \geq 0$ for all x

$\int_{-\infty}^{\infty} p(x)dx = 1$



$p(a \leq x \leq b) = \int_a^b p(x)dx$

Recall from last time: The Gaussian distribution

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

$$= Ce^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

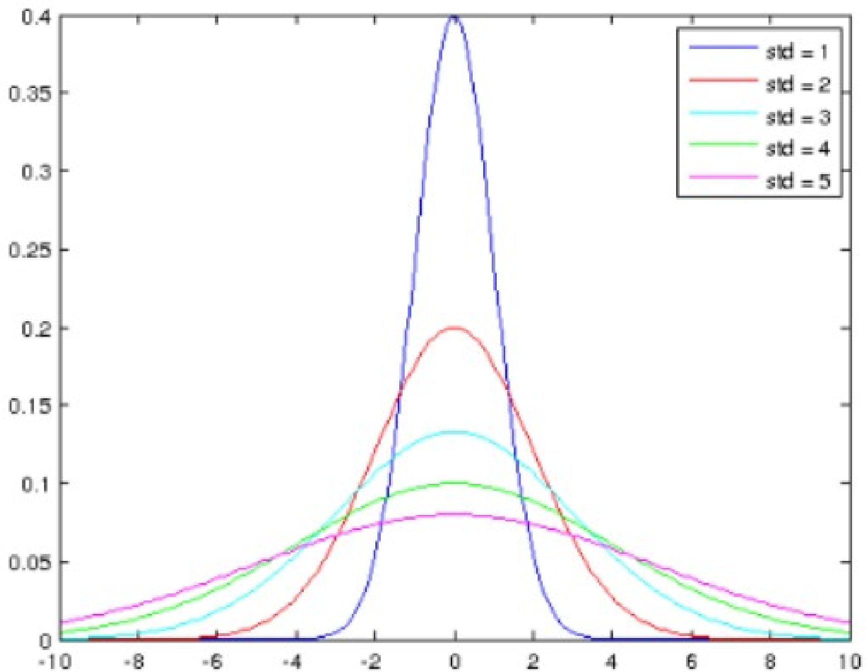
bandwidth maximum

μ is mean

σ^2 is variance

σ is standard deviation

$\beta = \frac{1}{\sigma^2}$ is precision



Multivariate Gaussian distribution

Multivariate Gaussian distribution

Multivariate Gaussian distribution

Probability Theory Arithmetic

	Discrete	**Continuous**
Variables	$X \in \{x_i\}_{i=1}^M, Y \in \{y_j\}_{j=1}^L$	$X \in \mathbb{R}, Y \in \mathbb{R}$
Example	$X = X_1$ eyes on first dice, $Y = X_1 + X_2$, sum of eyes	X =height of 4-year-old Y =height of mother
Sum rule	$p(x_i) = p(X = x_i) = \sum_j p(x_i, y_j)$	$p(x_i) = p(X = x_i) = \int p(x, y) dy$
Product rule	$p(X, Y) = p(Y\|X)p(X)$	$p(x, y) = p(y\|x)p(x)$

Probability Theory Arithmetic

	Discrete	Continuous
Variables	$X \in \{x_i\}_{i=1}^{M}, Y \in \{y_j\}_{j=1}^{L}$	$X \in \mathbb{R}, Y \in \mathbb{R}$
Example	$X = X_1$ eyes on first dice, $Y = X_1 + X_2$, sum of eyes	X =height of 4-year-old Y =height of mother
Sum rule	$p(x_i) = p(X = x_i) = \sum_j p(x_i, y_j)$	$p(x_i) = p(X = x_i) = \int p(x, y) dy$
Product rule	$p(X, Y) = p(Y\|X)p(X)$	$p(x, y) = p(y\|x)p(x)$
Independence	$p(X, Y) = p(X)p(Y)$	$p(x, y) = p(x)p(y)$

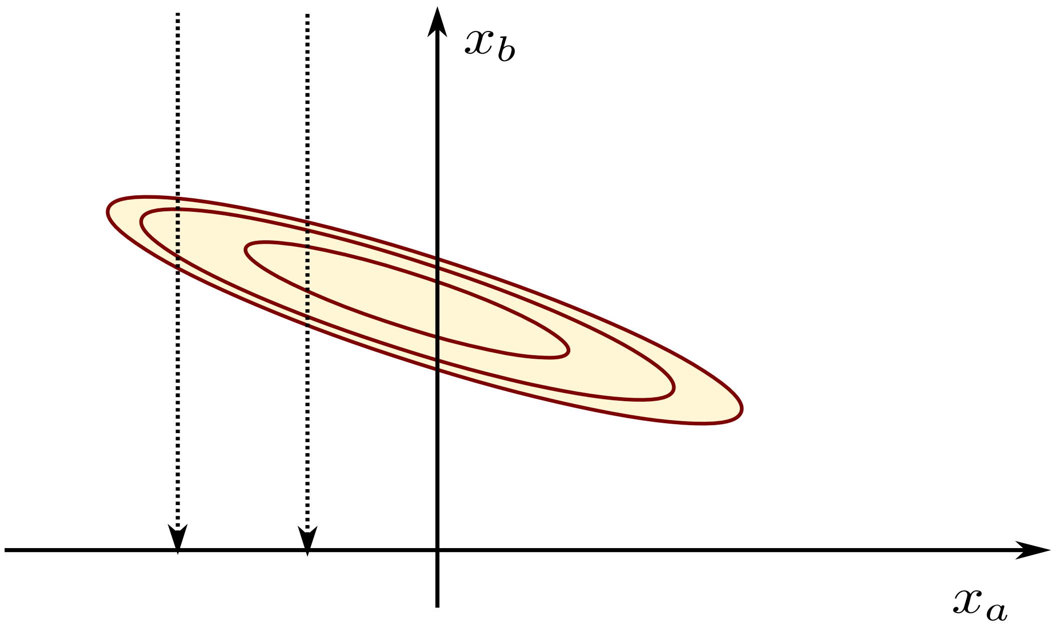
Exercise: $\boxed{p(y|x) = p(y) \text{ if } x \text{ and } y \text{ are independent}}$

Example: Marginal of Gaussian

Let the joint probability $p(\mathbf{x}_a, \mathbf{x}_b)$ be Gaussian

The marginal $p(\mathbf{x}_a)$ can be estimated with the sum rule

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$$



You will meet this in Assignment 1.3.1!

Example: Conditional of Gaussian

Let the joint probability $p(\mathbf{x}_a, \mathbf{x}_b)$ be Gaussian

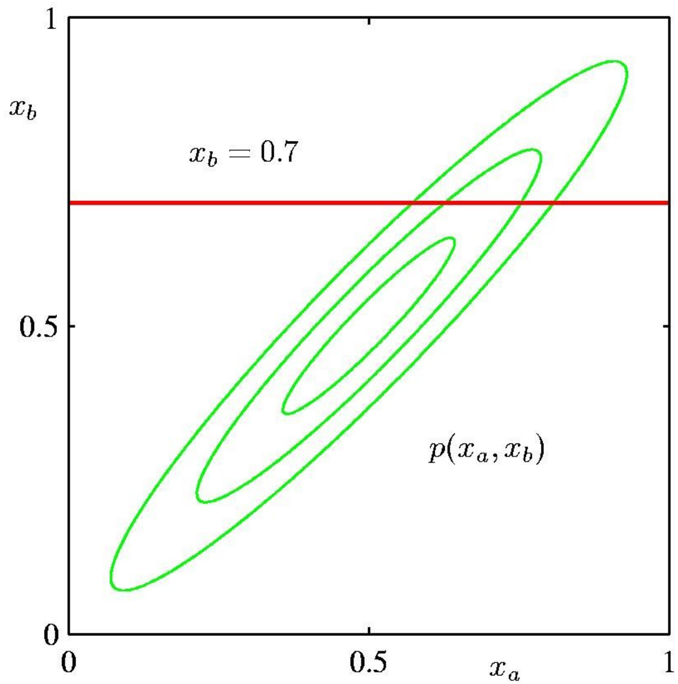
$$x = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix} \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix}$$

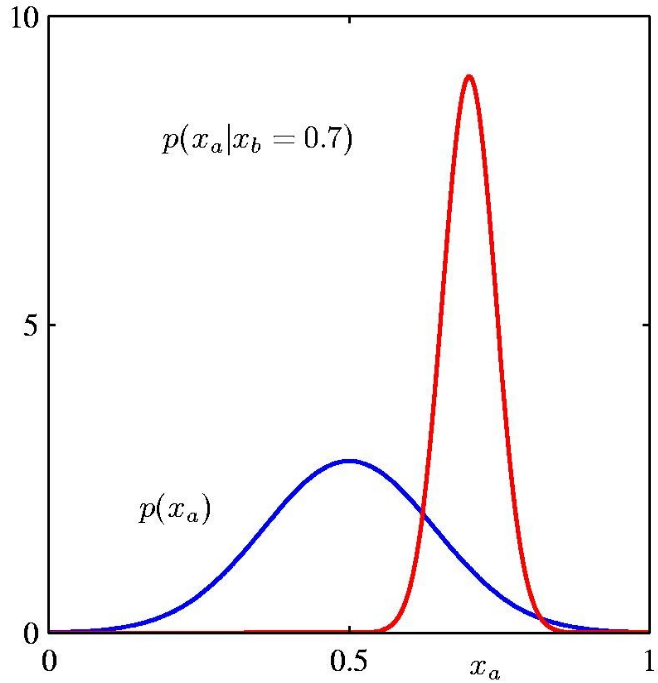
Express conditional using product rule $p(\mathbf{x}_a, \mathbf{x}_b) = p(\mathbf{x}_a|\mathbf{x}_b)p(\mathbf{x}_b)$

$$\boxed{\Lambda_{ij} \neq \Sigma_{ij}^{-1}}$$

The conditional $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\mu_{a|b}, \Lambda_{aa}^{-1})$

Meet this in Assignment I.3.1





Completing the square

Trick! Any function

$$f(x) = Ce^{c_1 x^2 + c_2 x + c_3}$$

can be normalized to become a Gaussian probability density function

$$\mathcal{N}(x|\mu, \sigma^2) = \tilde{C}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

because

$$c_1 x^2 + c_2 x + c_3 = c_1\left(x + \frac{c_2}{2c_1}\right)^2 + \left(c_3 + \frac{c_2^2}{4c_1}\right)$$

so

$$f(x) = Ce^{c_3 + \frac{c_2^2}{4c_1}} e^{c_1\left(x - \frac{-c_2}{2c_1}\right)^2}$$

To find the μ, σ

$$-\frac{1}{2\sigma^2}(x-\mu)^2 = -\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) = -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} + const.$$

and set

$$c_1 = -\frac{1}{2\sigma^2}, c_2 = \frac{\mu}{\sigma^2}$$

Similarly for multidimensional distributions -- Assignment I.3

Expectation Values

weighted average of statistic or function f over distribution

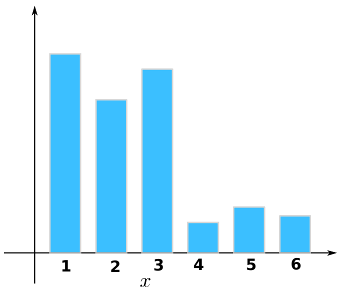
$\mathbb{E}[f(X)] = \sum_x f(x)p(x)$ **discrete**

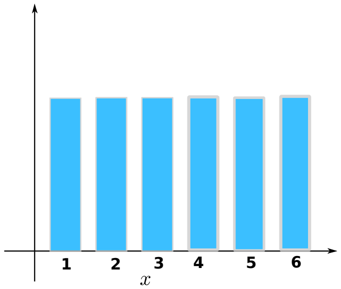
$\mathbb{E}[f(X)] = \int f(x)p(x)dx$ **continuous**

Examples:

mean of X $\mathbb{E}[X] =$ $\sum_x xp(x)$ **discrete**

$\int xp(x)dx$ **continuous**





Expectation Values

weighted average of statistic or function f over distribution

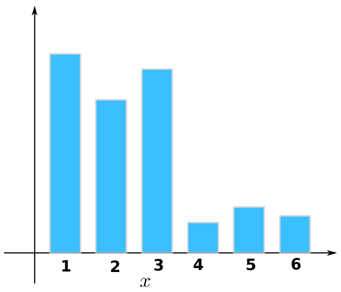
$\mathbb{E}[f(X)] = \sum_x f(x)p(x)$ **discrete**

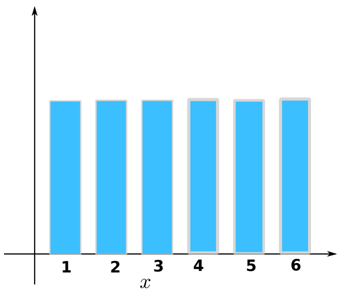
$\mathbb{E}[f(X)] = \int f(x)p(x)dx$ **continuous**

Examples:

mean of X $\mathbb{E}[X] =$ $\sum_x xp(x)$ **discrete**

$\int xp(x)dx$ **continuous**





variance of X $var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$

covariance of X $cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Multivariate Gaussian distribution: Covariance!

Covariance matrix!

Multivariate Gaussian distribution: Covariance!

Covariance matrix!

Multivariate Gaussian distribution: Covariance!

Covariance matrix!

Expectation Values

weighted average of statistic or function f over distribution

$\mathbb{E}[f(X)] = \sum_x f(x)p(x)$ **discrete**

$\mathbb{E}[f(X)] = \int f(x)p(x)dx$ **continuous**

Properties

$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

$\mathbb{E}[X+c] = \mathbb{E}[X] + c$

$\mathbb{E}[cX] = c\mathbb{E}[X]$

c is any constant

Exercise

- $var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- X, Y independent

  $\Rightarrow cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = 0$

Example: Gaussian

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

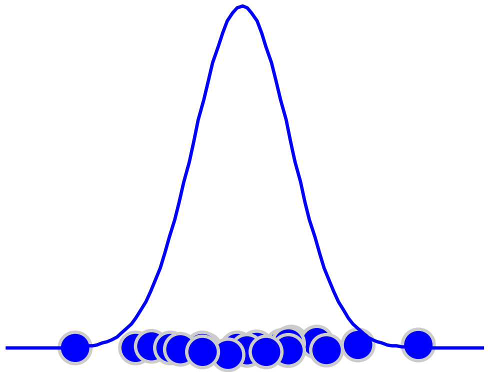
$$\mathbb{E}[f(X)] = \int f(x)p(x)dx$$

mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} xdx = \mu$

second moment: $\mathbb{E}[x^2] = \int_{\infty}^{\infty} \mathcal{N}(x|\mu, \sigma)x^2 = \mu^2 + \sigma^2$

variance: $var(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$

Today's problem: Fit distribution to dataset



Given:

N real-valued observations $X = x_1, \ldots, x_N$

Assume:

x_i drawn independently from *some* Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$

i.i.d. $=$ *independent* and *identically distributed*

Consequence of assumption:

$$\begin{aligned} p(X|\mu,\sigma^2) &= p(x_1, x_2, \ldots, x_N|\mu,\sigma^2) \\ &= p(x_1|\mu,\sigma^2)p(x_2|\mu,\sigma^2)\ldots p(x_N|\mu,\sigma^2) \\ &= \textstyle\prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2) \end{aligned}$$

a likelihood!

Task:

Find the Gaussian $\mathcal{N}(x|\mu, \sigma^2)$ which best fits the data

Strategy 1:

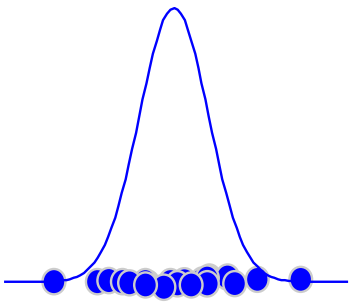
Find the Gaussian $\mathcal{N}(x|\mu_{ML}, \sigma^2_{ML})$ which maximises the likelihood

Today's problem: Fit distribution to dataset

Given:

N real-valued observations $X = x_1, \dots, x_N$

$p(X|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$ **a likelihood!**

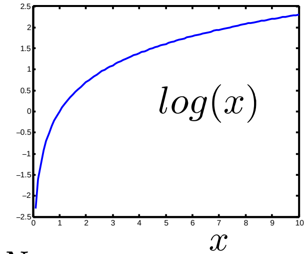


Strategy 1:

Find the Gaussian $\mathcal{N}(x|\mu_{ML},\sigma^2_{ML})$ which maximises the likelihood

Obs!

$*$ Dataset is fixed
$*$ Variables for fitting/optimizing are μ and σ^2



Trick!

Math problem: maximize

$$\log p(X|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi)$$

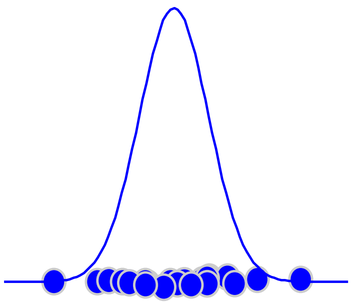
How to do that?

Today's problem: Fit distribution to dataset

Given:

N real-valued observations $X = x_1, \ldots, x_N$

$p(X|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$ **a likelihood!**

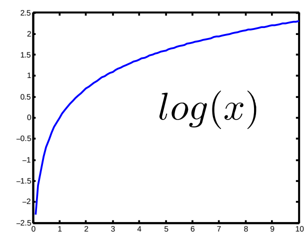


Strategy 1:

Find the Gaussian $\mathcal{N}(x|\mu_{ML}, \sigma^2_{ML})$ which maximises the likelihood

Obs!

$*$ Dataset is fixed
$*$ Variables for fitting/optimizing are μ and σ^2



Math problem: maximize

$$\log p(X|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi)$$

How to do that?

$$\frac{\partial}{\partial \mu} \log p(X|\mu, \sigma^2) = 0$$

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

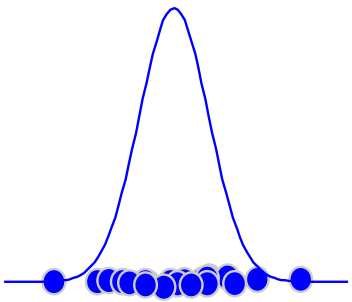
look familiar?

Today's problem: Fit distribution to dataset

Given:

N real-valued observations $X = x_1, \ldots, x_N$

$p(X|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$ **a likelihood!**

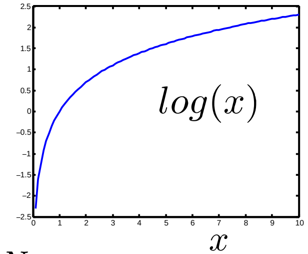


Strategy 1:

Find the Gaussian $\mathcal{N}(x|\mu_{ML}, \sigma^2_{ML})$ which maximises the likelihood

Obs!

$*$ Dataset is fixed
$*$ Variables for fitting/optimizing are μ and σ^2



Math problem: maximize

$$\log p(X|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi)$$

How to do that?

$$\frac{\partial}{\partial a} \log p(X|\mu_{ML}, a), \quad \sigma^2 = a$$

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

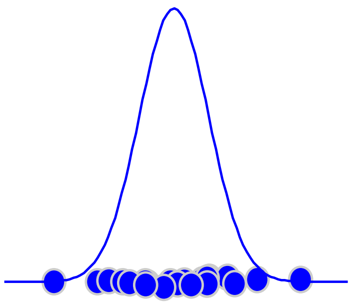
look familiar?

Today's problem: Fit distribution to dataset

Given:

N real-valued observations $X = x_1, \ldots, x_N$

$p(X|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$ **a likelihood!**



Strategy 1:

Find the Gaussian $\mathcal{N}(x|\mu_{ML}, \sigma^2_{ML})$ which maximises the likelihood

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

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

For multivariate Gaussians:

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$$

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \mu_{\mathbf{ML}})(\mathbf{x}_n - \mu_{\mathbf{ML}})^T$$

Second approach to parameter estimation

Bayesian statistics

Bayes' Rule

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

Proof? (Hint: the product rule) $p(x, y) = p(x|y)p(y)$

Why is this a useful theorem? Remember our task of the day!

N real-valued observations $X = x_1, \ldots, x_N$

x_i drawn independently from *some* Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$

i.i.d. = *independent* and *identically distributed*

Task:

Find the Gaussian $\mathcal{N}(x|\mu, \sigma^2)$ which best fits the data

Strategy 2:

Find parameters (μ, σ^2) such that $\mathcal{N}(x|\mu, \sigma^2)$ agrees the most with X.

Maximise $p\left((\mu, \sigma^2)|X\right)$

Bayes' Rule

$$\boxed{p(y|x) = \frac{p(x|y)p(y)}{p(x)}}$$

General setting:

w = model parameters

D = observed data

$p(D|w)$ = likelihood of data given model

$p(w|D)$ = probability distribution of model given data

Likelihood -- we get this from the model and the data

$$= \frac{p(D|w)p(w)}{p(D)}$$

Prior knowledge

Evidence -- a constant when the data is fixed

$$\rightsquigarrow p(w|D) \propto p(D|w)p(w)$$

Choosing Priors – conjugate priors

Optimize for computability:
Choose prior which multiplies **nicely** with likelihood
(that is, which leads to an algebraically nice analytic expression)
Called a **conjugate** prior!

Probability density distribution of model given data

Likelihood -- we get this from the model and the data

$$p(w|D) \propto p(D|w)p(w)$$

Prior knowledge and/or conjugate prior

Remember: Product of Gaussians is (unnormalized) Gaussian

Maximum a Posteriori (MAP) estimate

Example: Infer mean μ, assume known variance σ^2

Dataset $X = \{x_1, \ldots, x_N\}$ as before

Likelihood:

$p(X|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$

Prior:

$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$

Maximum a Posteriori (MAP) estimate

Example: Infer mean μ, assume known variance σ^2

Likelihood:

$$p(X|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

Prior:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Distribution we want to estimate:

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)\mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Gaussian! (complete the square)

$$= \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

Completing the square

Trick! Any function

$$f(x) = Ce^{c_1 x^2 + c_2 x + c_3}$$

can be normalized to become a Gaussian probability density function

$$\mathcal{N}(x|\mu, \sigma^2) = \tilde{C}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

because

$$c_1 x^2 + c_2 x + c_3 = c_1(x + \frac{c_2}{2c_1})^2 + (c_3 - \frac{c_2^2}{4c_1})$$

so

$$f(x) = Ce^{c_3 + \frac{c_2^2}{4c_1}} e^{c_1(x - \frac{-c_2}{2c_1})^2}$$

To find the μ, σ

$$-\frac{1}{2\sigma^2}(x-\mu)^2 = -\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) = -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} + const.$$

and set

$$c_1 = -\frac{1}{2\sigma^2}, c_2 = \frac{\mu}{\sigma^2}$$

Similarly for multidimensional distributions -- Assignment I.3

Maximum a Posteriori (MAP) estimate

Example: Infer mean μ, assume known variance σ^2

Likelihood:

$$p(X|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

Prior:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Distribution we want to estimate:

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)\mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Gaussian! (complete the square)

$$= \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$= \prod_{n=1}^{N} C_n e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} C_\mu e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

Maximum a Posteriori (MAP) estimate

Example: Infer mean μ, assume known variance σ^2

Likelihood:

$$p(X|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

Prior:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Distribution we want to estimate:

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)\mathcal{N}(\mu|\mu_0, \sigma_0)^2$$

Gaussian! (complete the square)

$$= \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$= \prod_{n=1}^{N} C_n e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} C_\mu e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

so the exponential is $-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2$

$$= \mu^2(-\frac{N}{2\sigma^2} - \frac{1}{2\sigma_0^2}) + \mu(-\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - \frac{\mu_0}{\sigma_0^2}) + const.$$

Maximum a Posteriori (MAP) estimate

Example: Infer mean μ, assume known variance σ^2

Likelihood:

$$p(X|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

Prior:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Distribution we want to estimate:

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)\mathcal{N}(\mu|\mu_0, {\sigma_0}^2)$$

Gaussian! (complete the square)

$$= \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$= \prod_{n=1}^{N} C_n e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} C_\mu e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

so the exponential is $-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2$

$$= \underbrace{\mu^2(-\frac{N}{2\sigma^2} - \frac{1}{2\sigma_0^2})}_{c_1} + \mu\underbrace{(-\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - \frac{\mu_0}{\sigma_0^2})}_{c_2} + const.$$

in the polynomial $c_1\mu^2 + c_2\mu + c_3$

Completing the square

Trick! Any function

$$f(x) = Ce^{c_1 x^2 + c_2 x + c_3}$$

can be normalized to become a Gaussian probability density function

$$\mathcal{N}(x|\mu, \sigma^2) = \tilde{C}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

because

$$c_1 x^2 + c_2 x + c_3 = c_1(x + \frac{c_2}{2c_1})^2 + (c_3 - \frac{c_2^2}{4c_1})$$

so

$$f(x) = Ce^{c_3 + \frac{c_2^2}{4c_1}} e^{c_1(x - \frac{-c_2}{2c_1})^2}$$

To find the μ, σ

$$-\frac{1}{2\sigma^2}(x-\mu)^2 = -\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) = -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} + const.$$

and set

$$c_1 = -\frac{1}{2\sigma^2}, c_2 = \frac{\mu}{\sigma^2}$$

Similarly for multidimensional distributions -- Assignment I.3

Maximum a Posteriori (MAP) estimate

Example: Infer mean μ, assume known variance σ^2

Likelihood:

$$p(X|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

Prior:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Distribution we want to estimate:

$$p(\mu|X) \propto p(X|\mu)p(\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)\mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Gaussian!
(complete the square)

$$= \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$= \prod_{n=1}^{N} C_n e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} C_\mu e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

so the exponential is $-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2$

$$= \underbrace{\mu^2\left(-\frac{N}{2\sigma^2} - \frac{1}{2\sigma_0^2}\right)}_{c_1} + \mu\underbrace{\left(-\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - \frac{\mu_0}{\sigma_0^2}\right)}_{c_2} + const.$$

in the polynomial $c_1\mu^2 + c_2\mu + c_3$

and we compute

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0}{N\sigma_0^2+\sigma^2}\mu_{ML} \qquad \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\mu_{MAP} = \text{argmax}_\mu p(\mu|X) = \mu_N$$

Bayesian Statistics – ML lingo

The formula $$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

In plain words $$p(model|data) = \frac{p(data|model)p(model)}{p(data)}$$

ML lingo $$posterior = \frac{likelihood \cdot prior}{evidence}$$

Two ways of estimating the distribution

Task:

Find the Gaussian $\mathcal{N}(x|\mu, \sigma^2)$ which best fits the data

Maximum Likelihood (ML)= estimate:

Choose parameters w that maximize $p(D|w)$

(likelihood function)

Maximum a Posteriori (MAP) estimate

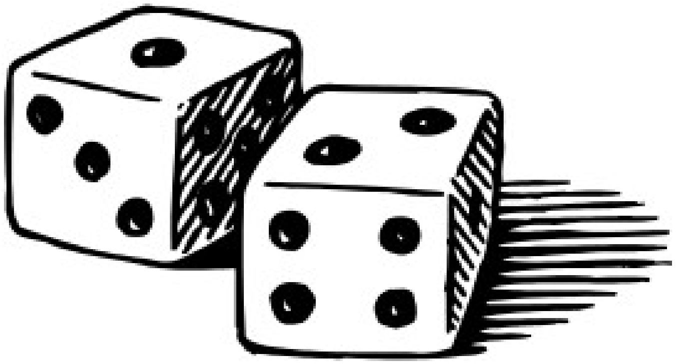
Choose parameters w that maximize $p(w|D)$

(posterior probability)

Non-parametric estimation

- Sometimes, it is not possible to model a probability density function parametrically
- Estimate it from the data!

- Unfair dice



- Image patches

Histogram

- A histogram H(X) of the random variable X is a table of frequency counts of N experiments (or data points):
  - Subdivide the domain of X, e.g. the set of real numbers, into M bins of width Δ (bin volume in D-dim.).
  - 2. For the i'th bin, let H(i) be the frequency count of how many times X falls into the bin.
- Probability estimate: Probability of falling in the i'th bin

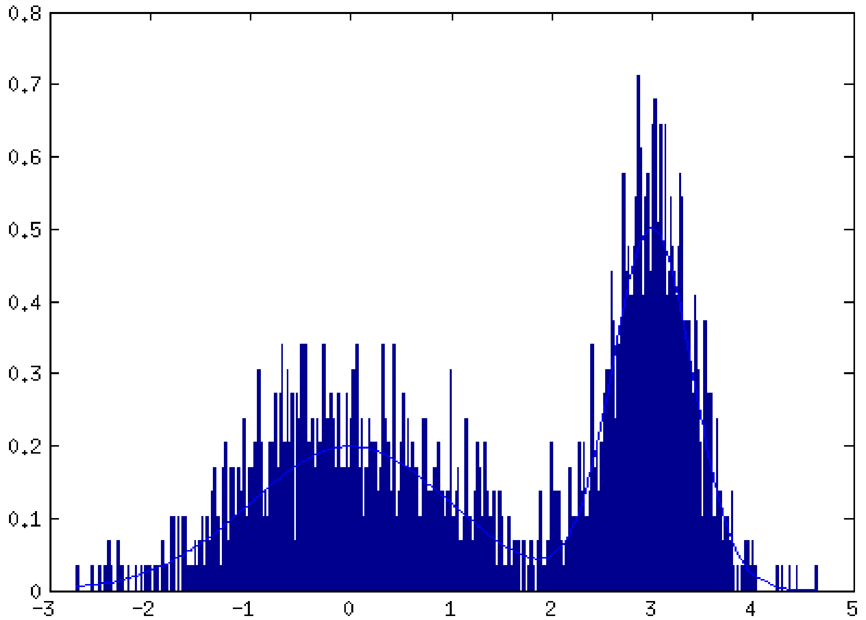
$$\mathrm{p}(\mathrm{X} \in \Delta i) = H(i)/N$$

(Probability estimator)

- Probability density estimate:

$$\mathrm{p}(\mathrm{x}) = \mathrm{H}(\mathrm{i})/(\mathrm{N}\ \Delta)$$

(Probability density estimator)



M = 500, N = 2000

Histogram

- A histogram H(X) of the random variable X is a table of frequency counts of N experiments (or data points):
  - Subdivide the domain of X, e.g. the set of real numbers, into M bins of width Δ (bin volume in D-dim.).
  - 2. For the i'th bin, let H(i) be the frequency count of how many times X falls into the bin.
- Probability estimate: Probability of falling in the i'th bin

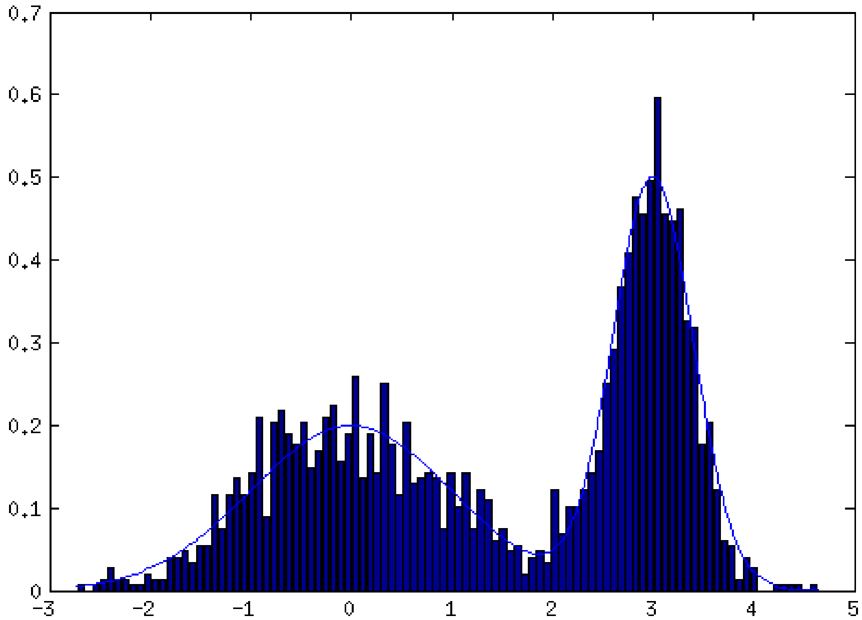
$$\mathrm{p}(\mathrm{X} \in \Delta i) = H(i)/N$$

(Probability estimator)

- Probability density estimate:

$$\mathrm{p}(\mathrm{x}) = \mathrm{H}(\mathrm{i})/(\mathrm{N}\ \Delta)$$

(Probability density estimator)



M = 100, N = 2000

Histogram

- A histogram H(X) of the random variable X is a table of frequency counts of N experiments (or data points):
  - Subdivide the domain of X, e.g. the set of real numbers, into M bins of width Δ (bin volume in D-dim.).
  - 2. For the i'th bin, let H(i) be the frequency count of how many times X falls into the bin.
- Probability estimate: Probability of falling in the i'th bin

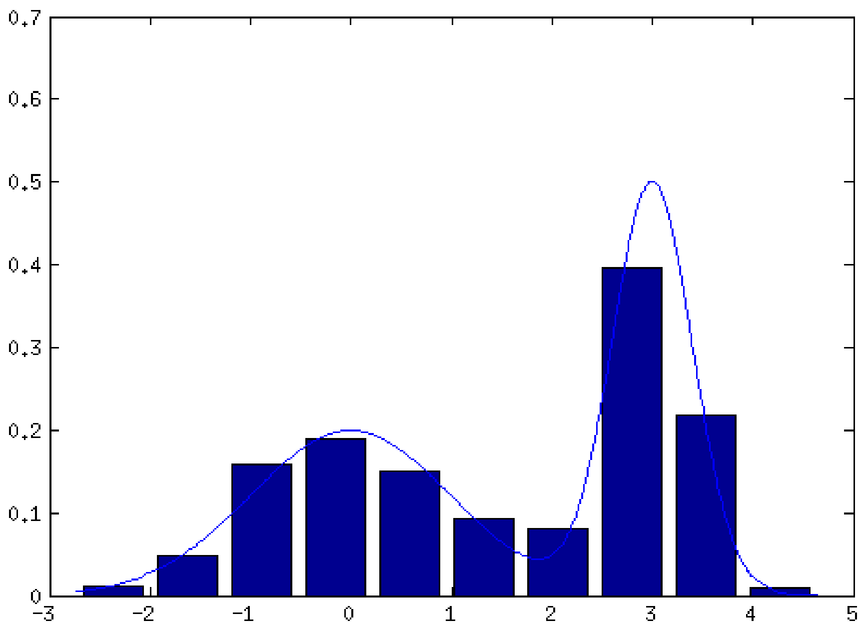
$$\mathrm{p}(\mathrm{X} \in \Delta i) = H(i)/N$$

(Probability estimator)

- Probability density estimate:

$$\mathrm{p}(\mathrm{x}) = \mathrm{H}(\mathrm{i})/(\mathrm{N}\ \Delta)$$

(Probability density estimator)



M = 10, N = 2000

Histogram

- A histogram H(X) of the random variable X is a table of frequency counts of N experiments (or data points):
  - Subdivide the domain of X, e.g. the set of real numbers, into M bins of width Δ (bin volume in D-dim.).
  - 2. For the i'th bin, let H(i) be the frequency count of how many times X falls into the bin.
- Probability estimate: Probability of falling in the i'th bin

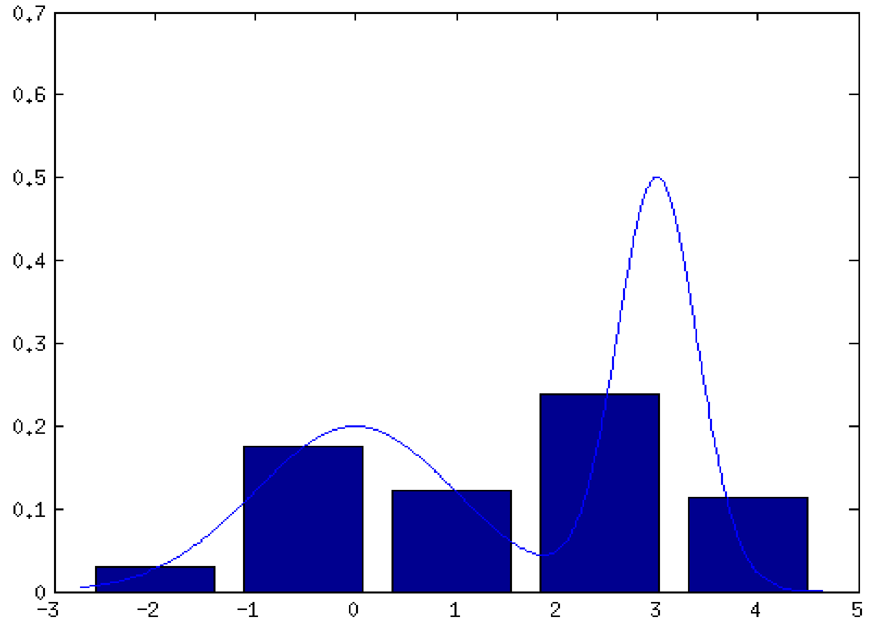
$$\mathrm{p}(\mathrm{X} \in \Delta i) = H(i)/N$$

(Probability estimator)

- Probability density estimate:

$$\mathrm{p}(\mathrm{x}) = \mathrm{H}(\mathrm{i})/(\mathrm{N}\ \Delta)$$

(Probability density estimator)



M = 5, N = 2000

Non-Parametric Density Estimation: Kernels (Parzen windows)

Replace histograms with estimates around arbitrary points $x$in $\mathbb{R}^D$

Count the number of points around x using a kernel function centered on x
kernel = bin)

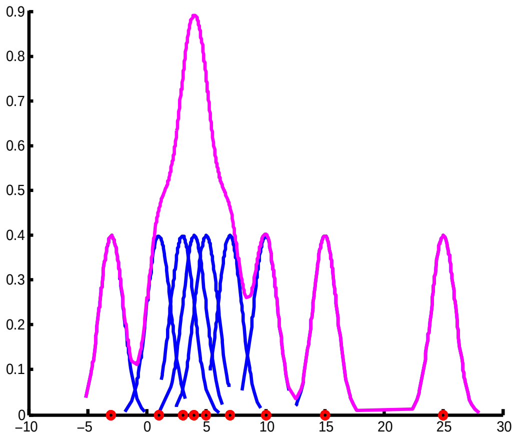
$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$

Equivalently, put a kernel centered on each data point x
and sum the values of the kernel functions at

Assume: The volume of the bin defined by the kernel is $V = h^D$

Probability density kernel estimate using a Gaussian kernel:

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{\sqrt{2\pi h}} e^{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}}$$



OBS! a "kernel" is not a "kernel" -- multiple uses of the word in ML

Why parametric estimation?

- **What's good about it?**
- **What's bad about it?**

Why parametric estimation?

- **What's good about it?**
  - Analytic expression
  - Computational speed
  - Precise solutions
- **What's bad about it?**
  - Restrictive choice of models

    (Gaussians are the nice, easy ones!)

Why nonparametric estimation?

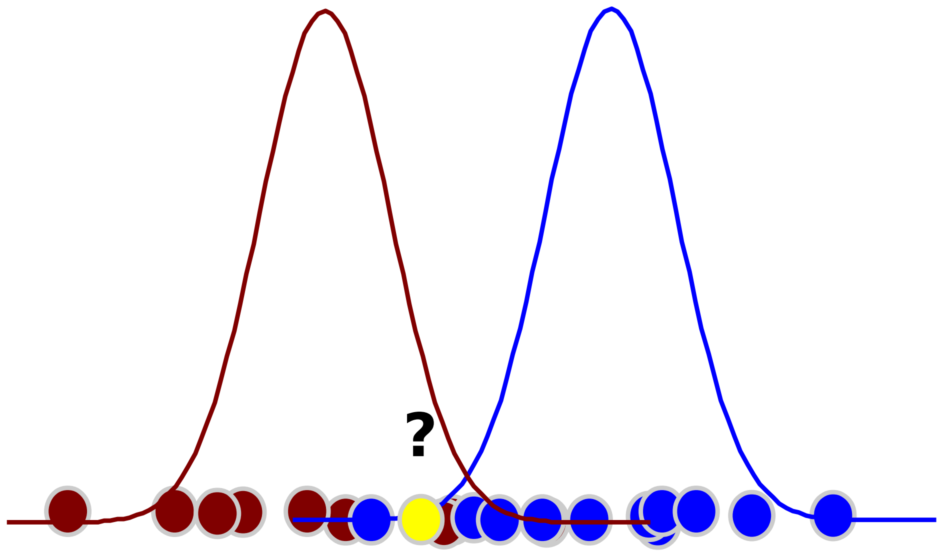
- **What's good about it?**

- **What's bad about it?**

Why nonparametric estimation?

- **What's good about it?**
  - No assumptions on distributions
  - Easy to understand and implement (the ones we've seen)
- **What's bad about it?**
  - Not exact
  - Computationally expensive

Recall: Probability distributions important for probabilistic ML...



- We will meet ML and MAP again in 2 weeks, for **regression**

Summary: After today's lecture you should

- Know the theoretical background for estimation of distributions
- Know the principles of Bayesian estimation
- Know standard techniques for parametric and non-parametric estimation of probability distributions
  - Maximum likelihood and maximum a posteriori estimation
  - Examples of non-parametric methods
  - Conjugate priors
- Be able to use the above parametric techniques for estimation of Gaussian distributions in real problems
- Corresponding reading material: (CB pages 1-28 and 78-113, 120-127)

Next time!

- Christian!
- Ingredients of statistical learning theory (loss, risk minimization, bounds)
- Reading material: (CB sections 1.3, 1.5, 2.5.2, 7.1.5; KBML sections 2.1 until 2.2.1)