University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 2 - Probability Theory - Maximum Likelihood - Maximum A Posteriori - Bayesian Prediction

Erik Bekkers

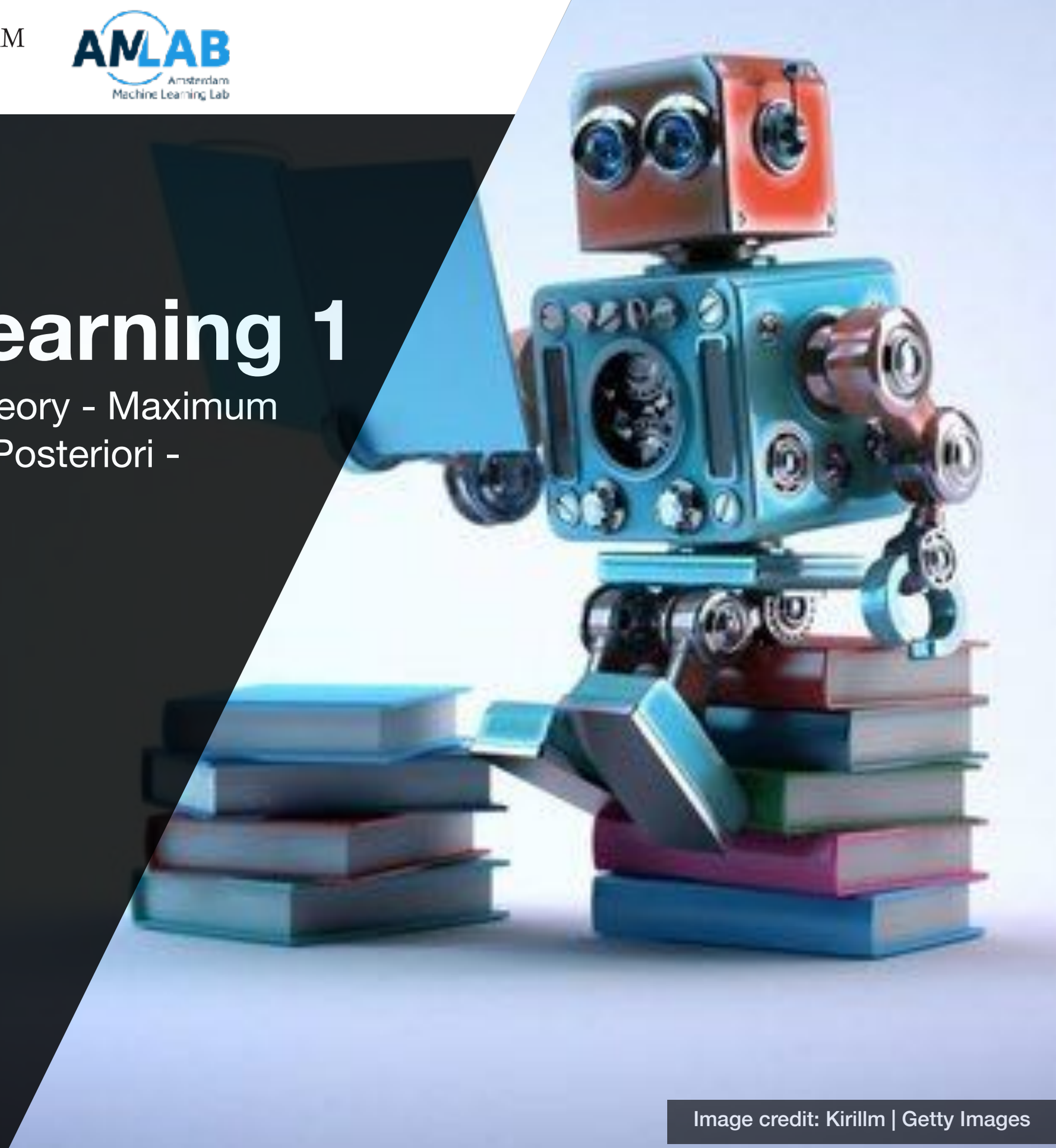


Image credit: Kirillm | Getty Images

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Recap basic probability + **independent random variables**

Erik Bekkers

Image credit: Kirillm | Getty Images

Slide credits: Patrick Forré and Rianne van den Berg

The Rules of Probability Theory

For random variables $x \in X$ and $y \in Y$:

	Discrete	Continous
Additivity	$p(x \in A) = \sum_{x \in A} p(X = x)$	$p(x \in (a, b)) = \int_a^b p(x)\mathrm{d}x$
Positivity	$p(X = x) \geq 0$	$p(X = x) \geq 0$
Normalization	$\sum_X p(X) = 1$	$\int_X p(x)\mathrm{d}x = 1$
Sum Rule	$p(X) = \sum_Y p(X, Y)$	$p(x) = \int_Y p(x, y)\mathrm{d}y$
Product Rule	$p(X, Y) = p(X \mid Y)p(Y)$	$p(x, y) = p(x \mid y)p(y)$

Bayes Theorem

- Product rule $p(x, y) = p(x \mid y)p(y)$
- Symmetry property $p(x, y) = p(y, x)$

$p(x|y)p(y) = p(y|x)p(x)$

- Bayes rule → $p(y \mid x) = \dfrac{p(x \mid y)p(y)}{p(x)}$

- Denominator: $\sum_{y \in Y} p(y \mid x) = 1$

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

- Formally: *X* and *Y* are called independent if

$$p(x, y) = p(x)\, p(y) \qquad \text{for all } x \in \mathcal{X},\ y \in \mathcal{Y}$$

- Equivalent to

$$p(x \mid y) = p(x)$$

$$\Rightarrow p(x, y) = p(x \mid y)\, p(y)$$

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 2.1 - Expectation - Variance

Erik Bekkers

(Bishop 1.2.2)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Expectations

- random variable $x \in X$ and function $f : X \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] = \begin{cases} \sum_{x \in X} f(x)\, p(x) & \text{discrete} \\ \int_X f(x)\, p(x)\, dx & \text{continuous} \end{cases}$$

"being explicit in what the expectation is taken over"

  - For *N* points drawn from p(X): $D = \{x_1, x_2, \ldots, x_N\}$

$$\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

- Conditional expectation:

$$\mathbb{E}[f \mid y] = \mathbb{E}_{x \sim p(X \mid Y = y)}[f(x)] = \begin{cases} \sum_{x \in X} f(x)\, p(x \mid y) \\ \int f(x)\, p(x \mid y)\, dx \end{cases}$$

Variance

- $\mathbb{E}[f(x)+g(x)] = \mathbb{E}[f(x)] + \mathbb{E}[g(x)]$
- $\mathbb{E}[c\,f(x)] = c\,\mathbb{E}[f(x)]$

linearity of $\mathbb{E}$

- $\mathbb{E}[c] = c$

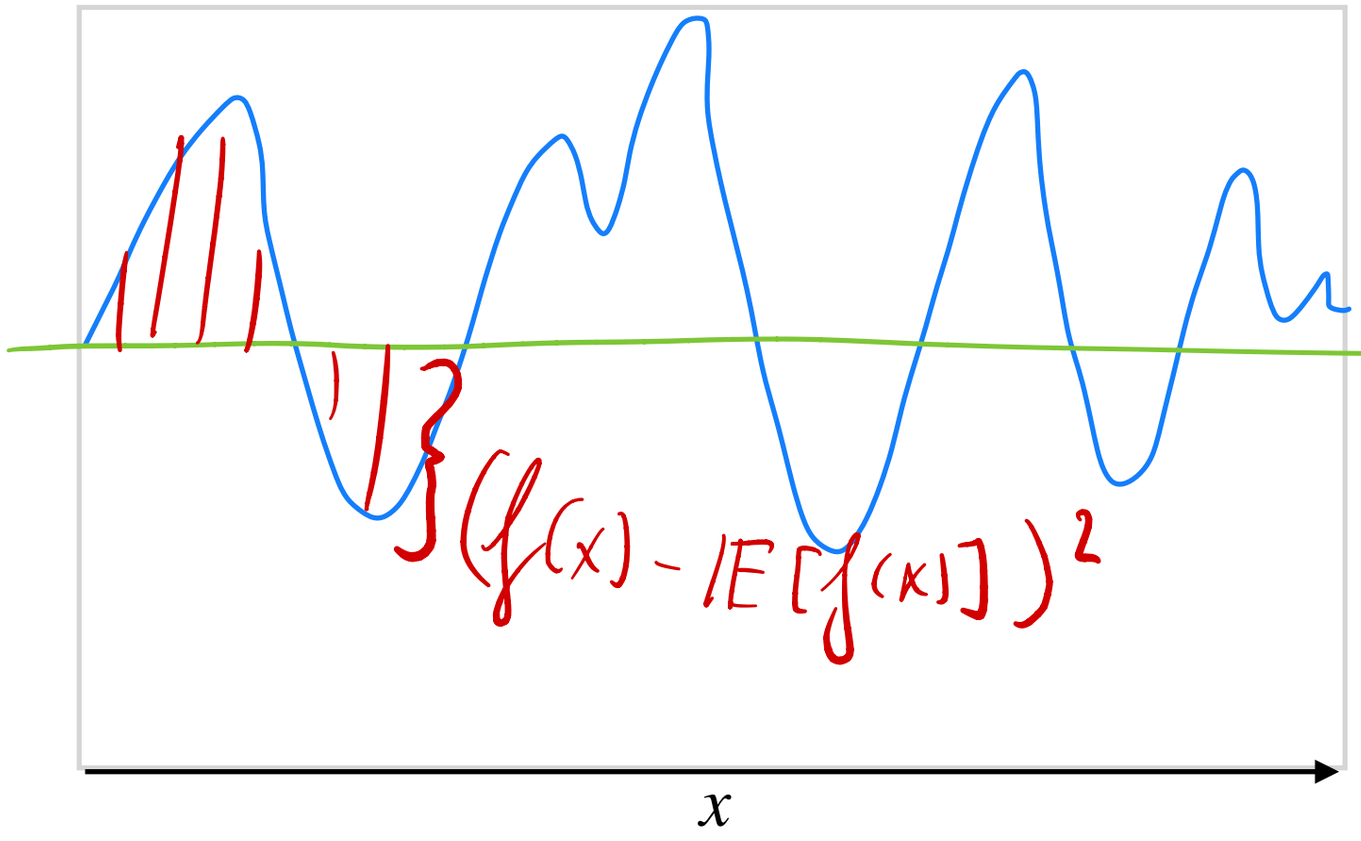
- The expected quadratic distance between f and its mean $\mathbb{E}[f]$

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

$$= \mathbb{E}[f(x)^2 - 2 f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2]$$

$$= \mathbb{E}[f(x)^2] - \mathbb{E}[2 f(x)\mathbb{E}[f(x)]] + \mathbb{E}[\mathbb{E}[f(x)]^2]$$

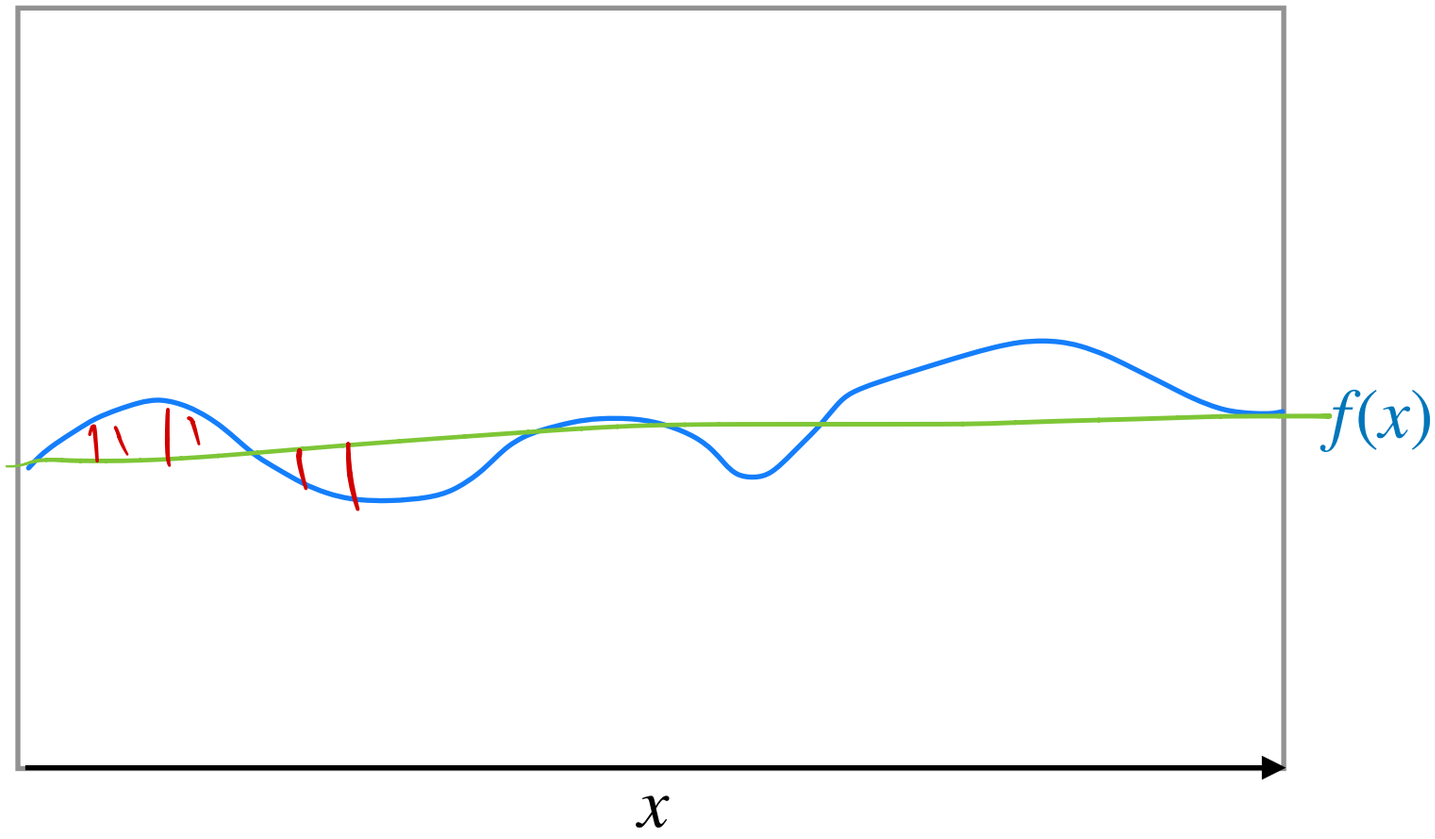
$$= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$



$\mathbb{E}[f(x)]$

$(f(x) - \mathbb{E}[f(x)])^2$

high variance



low variance

Covariance between 2 random variables

- Measures the extent to which *X* and *Y* vary together

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y\sim p(x,y)}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

$$= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

- Vectors of random variables $\mathbf{x}$ ($\in \mathbb{R}^D$) and $\mathbf{y}$, covariance matrix:

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x},\mathbf{y})}[\underbrace{(\mathbf{x} - \mathbb{E}[\mathbf{x}])}_{\mathbb{R}^{D\times 1}}\underbrace{(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T}_{\mathbb{R}^{1\times D}}] \in \mathbb{R}^{D\times D}$$

$$= \mathbb{E}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]^T$$

- Define $\mathrm{cov}[x] := \mathrm{cov}[x, x]$

Covariance between 2 random variables

- Covariance between independent variables

$$\mathrm{cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = 0$$

$$\int_X \int_Y x\, y\, p(x,y)\, dx\, dy = \int_X \int_Y x\, y\, p(x)\, dx\, p(y)\, dy = \int_X x\, p(x)\, dx \int_y y\, p(y)\, dy$$

$$= \mathbb{E}[x]\,\mathbb{E}[y]$$

- Note: $\mathrm{cov}[x, y] = 0$ does not imply x, y independent!

$$x \sim \underbrace{\mathcal{U}[-1, 1]}_{p(x) = \frac{1}{2}}$$

$$y = x^2$$

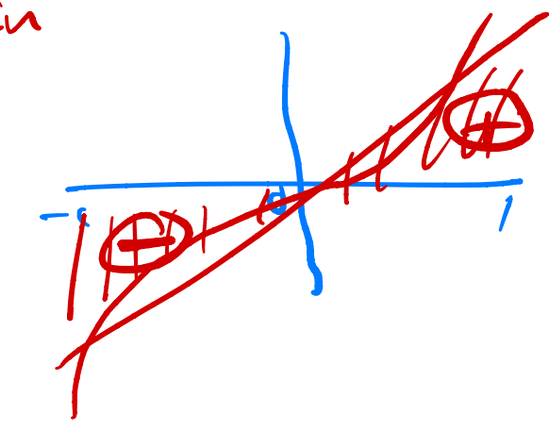
$$\mathbb{E}[xy] = \int x^3 p(x)\, dx = \int x^3 \tfrac{1}{2}\, dx = 0$$

$$\mathrm{cov}[x, y] = \underbrace{\mathbb{E}[xy]}_{0} - \underbrace{\mathbb{E}[x]}_{0}\mathbb{E}[y] = 0$$

x^2

x also odd function

because integrating odd function over symmetric domain



University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 2.2 - Gaussian Distribution

Erik Bekkers

(Bishop 1.2.4)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Gaussian Distribution

- The Normal/Gaussian distribution

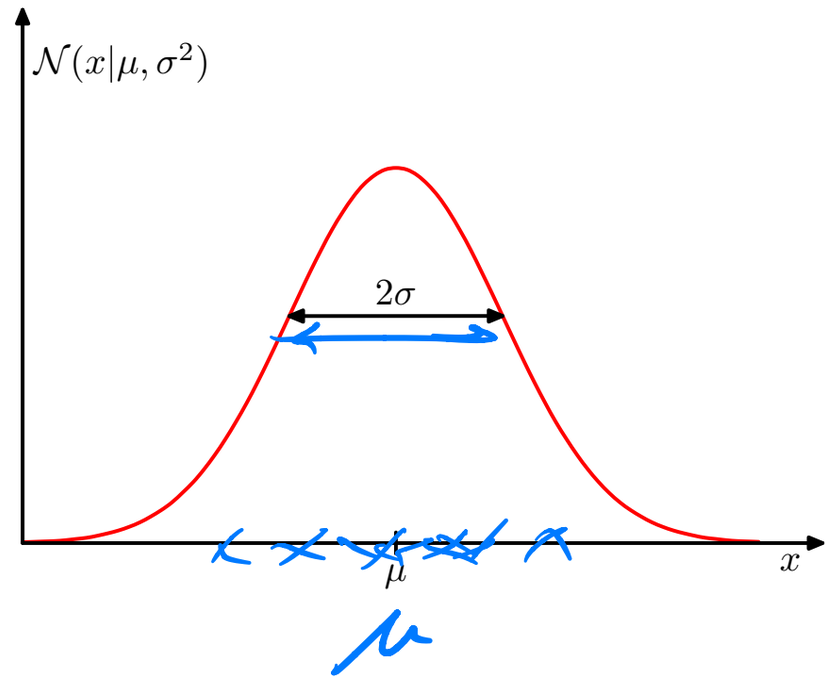
$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$\in \mathbb{R}$

mean

var : σ^2

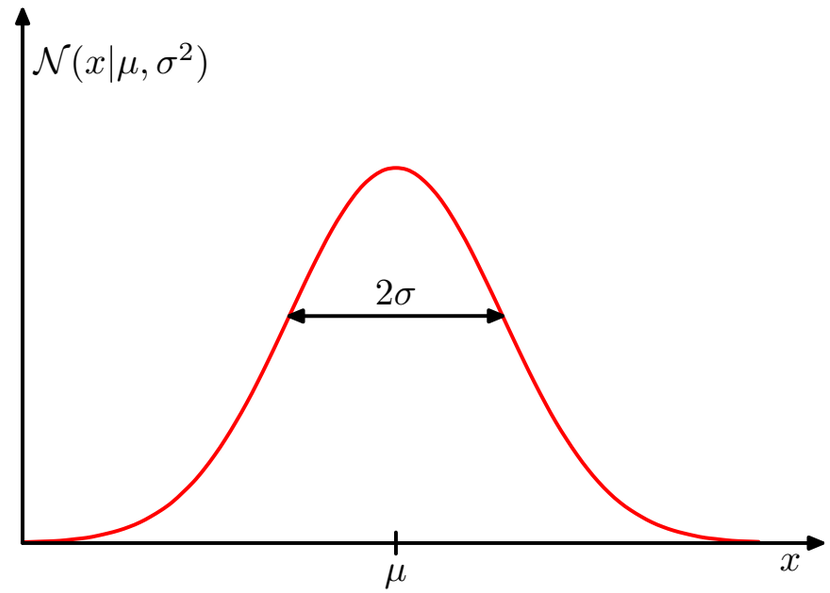
standard deviation : σ



Gaussian Distribution

- The Normal/Gaussian distribution

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



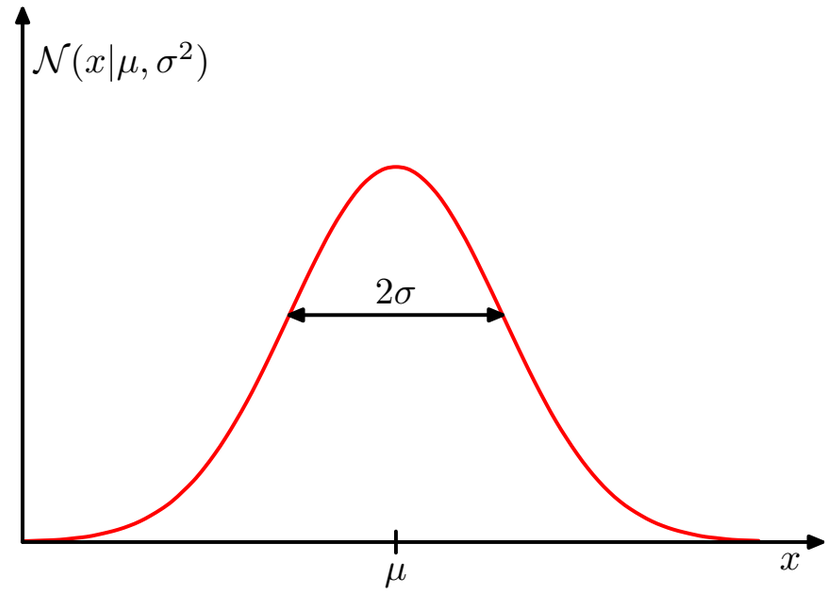
- **Mean** (of $x \sim \mathcal{N}(x\,|\,\mu,\sigma^2)$):

$$\mathbb{E}_{x\sim\mathcal{N}(x|\mu,\sigma^2)}[x] = \int_{-\infty}^{\infty} x\,\mathcal{N}(x\,|\,\mu,\sigma^2)\mathrm{d}x = \mu$$

Gaussian Distribution

- The Normal/Gaussian distribution

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



- **Mean** (of $x \sim \mathcal{N}(x\,|\,\mu,\sigma^2)$):

$$\mathbb{E}_{x\sim\mathcal{N}(x|\mu,\sigma^2)}[x] = \int_{-\infty}^{\infty} x\,\mathcal{N}(x\,|\,\mu,\sigma^2)\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\mathrm{d}x$$

simplify to e^{-y^2}

Step 1: Change of variables

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu) \quad \rightarrow \quad x = \sqrt{2\sigma^2}\,y + \mu$$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\sqrt{2\sigma^2}} \qquad \rightarrow \mathrm{d}x = \sqrt{2\sigma^2}\,\mathrm{d}y$$

2: Integration of odd funcs

$$\int_{-\infty}^{\infty} y\,e^{-y^2}\mathrm{d}y = 0$$

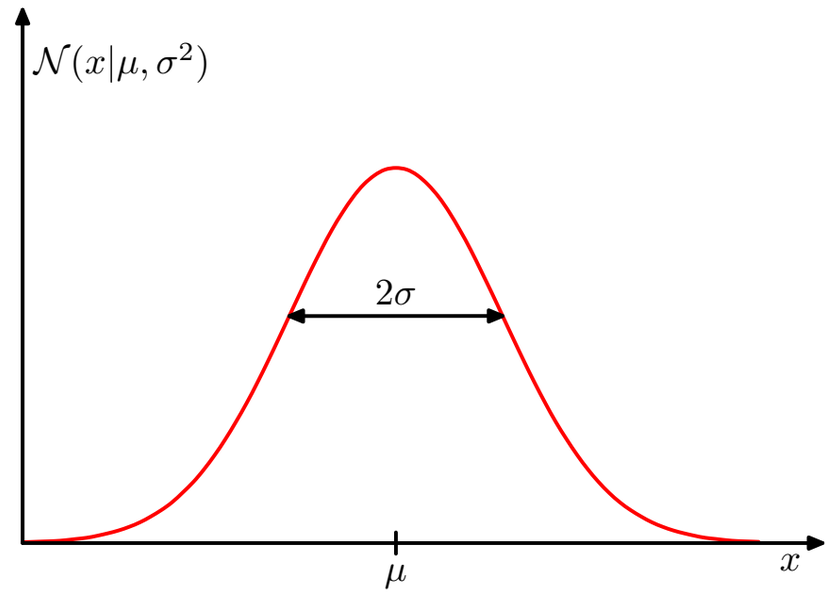
3: Useful property:

$$\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y = \sqrt{\pi}$$

Gaussian Distribution

- The Normal/Gaussian distribution

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



- **Mean** (of $x \sim \mathcal{N}(x\,|\,\mu,\sigma^2)$):

$$\mathbb{E}_{x\sim\mathcal{N}(x|\mu,\sigma^2)}[x] = \int_{-\infty}^{\infty} x\,\mathcal{N}(x\,|\,\mu,\sigma^2)\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} x\,\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\mathrm{d}x$$

$$\overset{(1)}{=} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}\,y+\mu)\,\frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2}\sqrt{2\sigma^2}\mathrm{d}y$$

Step 1: Change of variables

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu) \quad\rightarrow\quad x = \sqrt{2\sigma^2}\,y+\mu$$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\sqrt{2\sigma^2}} \quad\rightarrow \mathrm{d}x = \sqrt{2\sigma^2}\,\mathrm{d}y$$

2: Integration of odd funcs

$$\int_{-\infty}^{\infty} y\,e^{-y^2}\mathrm{d}y = 0$$

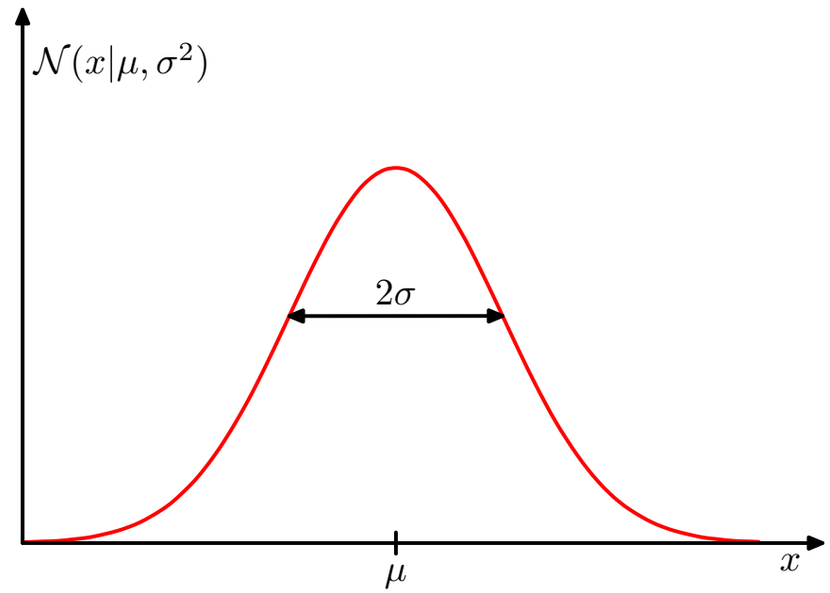
3: Useful property:

$$\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y = \sqrt{\pi}$$

Gaussian Distribution

- The Normal/Gaussian distribution

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



- **Mean** (of $x \sim \mathcal{N}(x\,|\,\mu,\sigma^2)$):

$$\mathbb{E}_{x\sim\mathcal{N}(x|\mu,\sigma^2)}[x] = \int_{-\infty}^{\infty} x\,\mathcal{N}(x\,|\,\mu,\sigma^2)\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} x\,\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\mathrm{d}x$$

$$\overset{(1)}{=} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}\,y + \mu)\,\frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2}\sqrt{2\sigma^2}\mathrm{d}y$$

$$= \int_{-\infty}^{\infty} \left(\sqrt{2\sigma^2}\,y + \mu\right)\,\frac{1}{\sqrt{\pi}} e^{-y^2}\mathrm{d}y$$

be odd function

Step 1: Change of variables

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu) \quad \rightarrow \quad x = \sqrt{2\sigma^2}\,y + \mu$$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\sqrt{2\sigma^2}} \quad \rightarrow \mathrm{d}x = \sqrt{2\sigma^2}\,\mathrm{d}y$$

2: Integration of odd funcs

$$\int_{-\infty}^{\infty} y\,e^{-y^2}\mathrm{d}y = 0$$

3: Useful property:

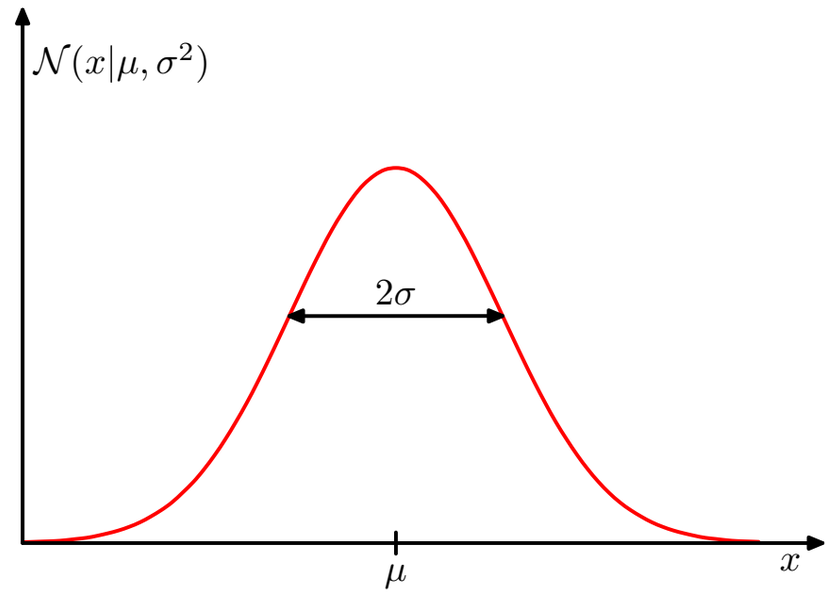
$$\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y = \sqrt{\pi}$$

can be derived from normalization property of distributions

Gaussian Distribution

- The Normal/Gaussian distribution

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



- **Mean** (of $x \sim \mathcal{N}(x\,|\,\mu,\sigma^2)$):

$$\mathbb{E}_{x\sim\mathcal{N}(x|\mu,\sigma^2)}[x] = \int_{-\infty}^{\infty} x\,\mathcal{N}(x\,|\,\mu,\sigma^2)\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} x\,\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\mathrm{d}x$$

$$\overset{(1)}{=} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}\,y+\mu)\,\frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2}\sqrt{2\sigma^2}\mathrm{d}y$$

$$= \int_{-\infty}^{\infty} \left(\sqrt{2\sigma^2}\,y\,+\,\mu\right)\,\frac{1}{\sqrt{\pi}} e^{-y^2}\mathrm{d}y \overset{(2+3)}{=} \mu$$

Step 1: Change of variables

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu) \quad\rightarrow\quad x = \sqrt{2\sigma^2}\,y+\mu$$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\sqrt{2\sigma^2}} \qquad\rightarrow \mathrm{d}x = \sqrt{2\sigma^2}\,\mathrm{d}y$$

2: Integration of odd funcs

$$\int_{-\infty}^{\infty} y\,e^{-y^2}\mathrm{d}y = 0$$

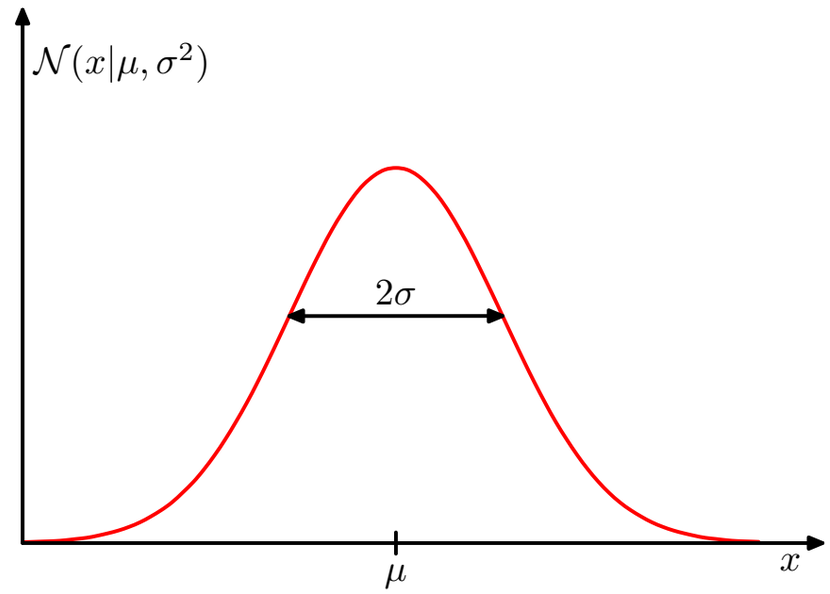
3: Useful property:

$$\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y = \sqrt{\pi}$$

Gaussian Distribution

- The Normal/Gaussian distribution

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



- **Variance** (of $x \sim \mathcal{N}(x\,|\,\mu,\sigma^2)$):

$$\mathrm{var}[x] = \int_{-\infty}^{\infty} (x-\mu)^2\, \mathcal{N}(x\,|\,\mu,\sigma^2)\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2\, \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\mathrm{d}x$$

… (**using** $1+2+4$)

$$= \sigma^2$$

see ch 2.3

Step 1: Change of variables

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu) \quad \rightarrow \quad x = \sqrt{2\sigma^2}\, y + \mu$$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\sqrt{2\sigma^2}} \qquad \rightarrow \mathrm{d}x = \sqrt{2\sigma^2}\, \mathrm{d}y$$

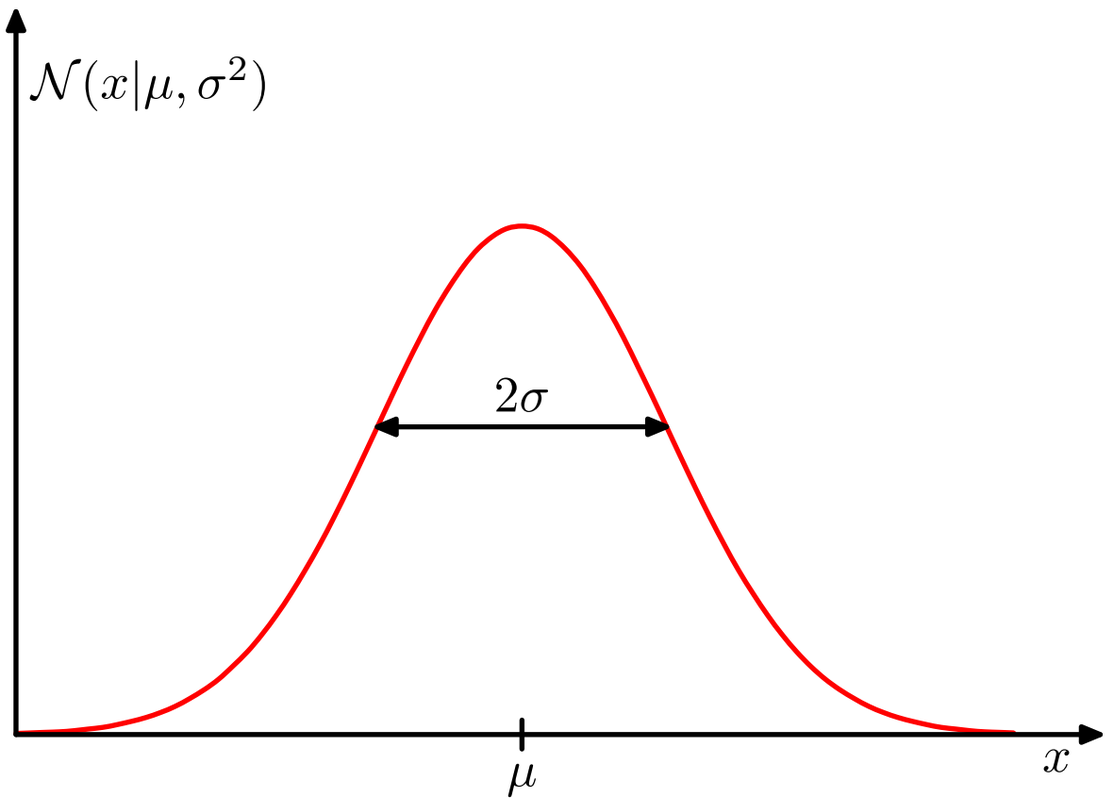
4: Another convenient trick:

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2}\mathrm{d}x = -\frac{\partial}{\partial a}\int_{-\infty}^{\infty} e^{-ax^2}\mathrm{d}x$$

$$= \frac{\partial}{\partial a}\sqrt{\frac{\pi}{a}} \quad = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}$$

Gaussian Distribution

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



$$x \sim \mathcal{N}(x\,|\,\mu,\sigma^2): \begin{cases} \mathbb{E}[x] = \mu \\ \mathrm{Var}[x] = \sigma^2 \end{cases}$$

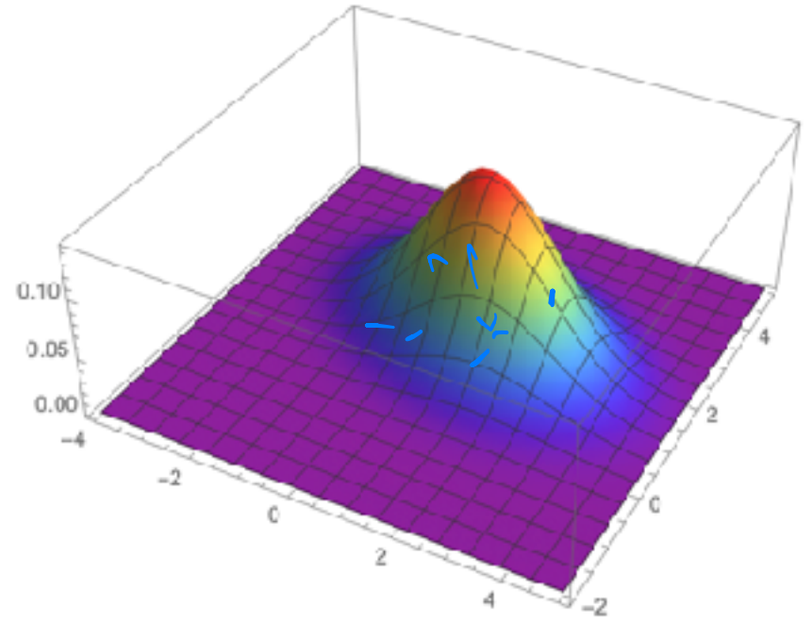
Multivariate Gaussian Distribution

- *D*-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_D)^T$

- $\mathcal{N}(\mathbf{x} \,|\, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D \,|\, \mathbf{\Sigma} \,|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu)}$ $\qquad |\,\mathbf{\Sigma}\,| = \det \mathbf{\Sigma}$

- $\mathbf{\Sigma} = Cov[\underline{x}, \underline{x}] \in \mathbb{R}^{D \times D}$

- $\mathbb{E}[\mathbf{x}] = \underline{\mu}$



1. Substitution $\mathbf{y} = (\mathbf{x} - \mu)$

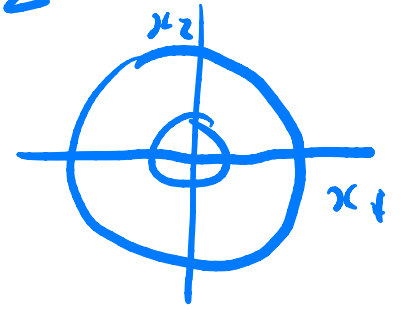
2. Normalization factor: $\int_{\mathbb{R}^D} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{A} \mathbf{x}} = \frac{2\pi^{D/2}}{|\,\mathbf{A}\,|^{1/2}}$ *(since* $\int_{\mathbb{R}^D} \mathcal{N}(\mathbf{x} \,|\, \mu, \mathbf{\Sigma}) = 1$*)*

$N(x \mid \mu, \Sigma)$

anisotropic

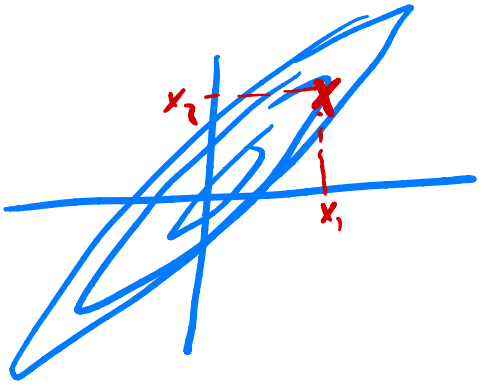
isotropic

$\Sigma = I = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$



x_1 and x_2 are not correlated

extremely anisotropic



x_1 and x_2 are highly correlated

knowing x_1 is knowing x_2

University of Amsterdam
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

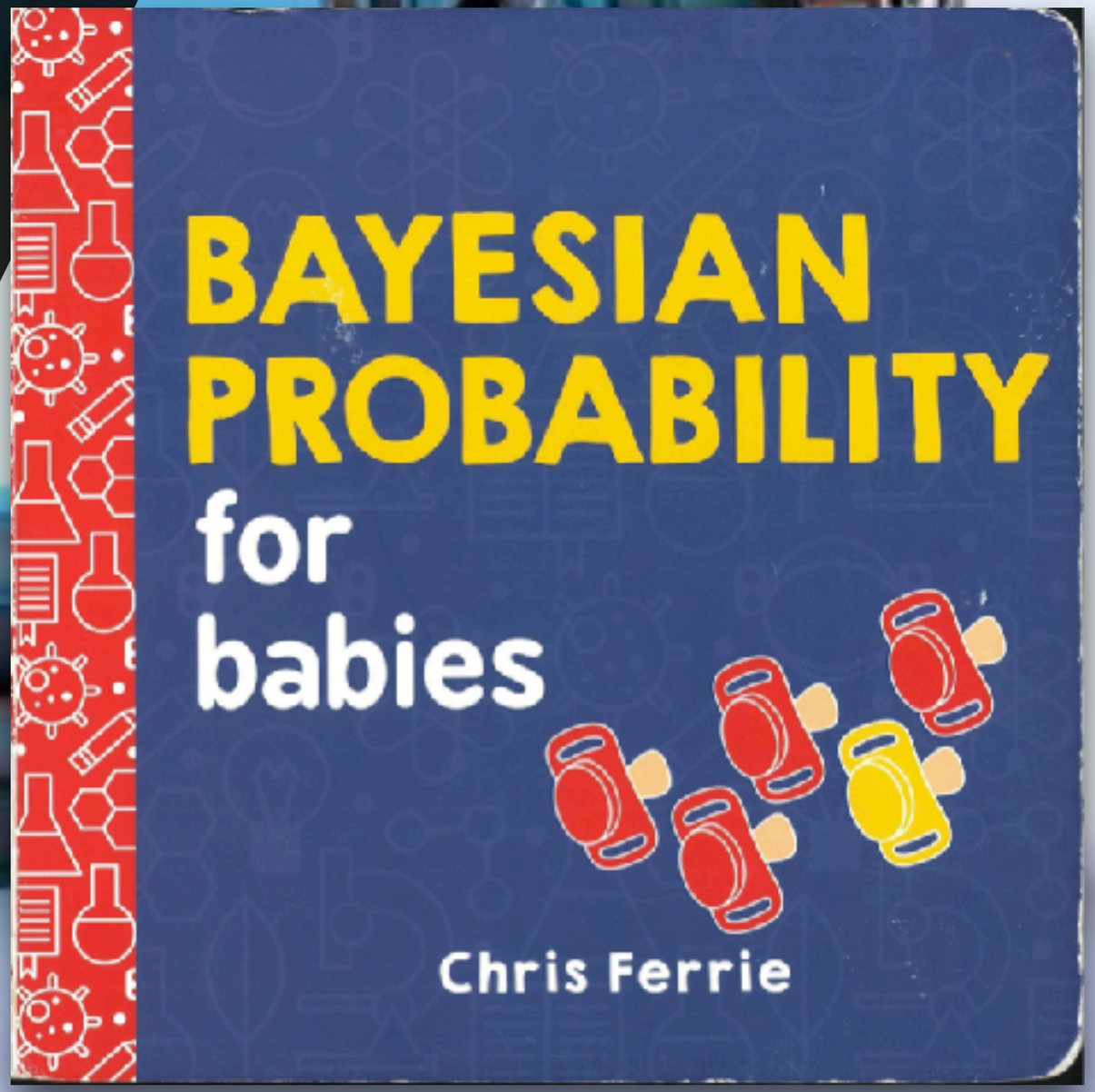
Machine Learning 1

Example Bayes Rule

Erik Bekkers

(Bishop 1.2.0 - 1.2.1)

Getting into model estimation



Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

This is a ball.

The ball is a piece of candy on a cookie. Yum!

Some cookies have candy.

Some don't.

Cookie

$C = \{ \text{candy}, \text{no-candy} \}$

$\{ 1, 0 \}$

B = 0

Take a bite. It has no candy.

C = 1

B = 0

Did it come from a candy cookie?

$B = \{1, 0\}$

$= \{\text{with candy}, \text{no candy}\}$

Either it came from a candy cookie,

or it didn't. What are the chances?

likelihood empty bgtr came from candy cookie

$p(B=0 \mid C=1)$

$p(B=0 \mid C=0)$

posterior prob

$p(C=1 \mid B=0)$

$p(C=0 \mid B=0)$

If the cookie had no candy,
then every bite would have no candy.

The probability of a no-candy bite,
given a no-candy cookie, is 1.

$p(B=0 \mid C=0) = 1$

"likelihood under the no candy model"

Pr(|)

=

1/3

If the cookie had candy, then very few bites would have no candy.

The probability of a no-candy bite, given a candy cookie, is 1/3.

$p(B=G \mid C=1) = \frac{1}{3}$

Pr(|) > Pr(|)

1 is greater than 1/3.

So the no-candy bite ~~probably~~ likely came from a no-candy cookie!

But what if we knew there were 10 cookies, and all had candy but one?

prior $p(C=1) = \frac{9}{10}$

$p(C=0) = \frac{1}{10}$

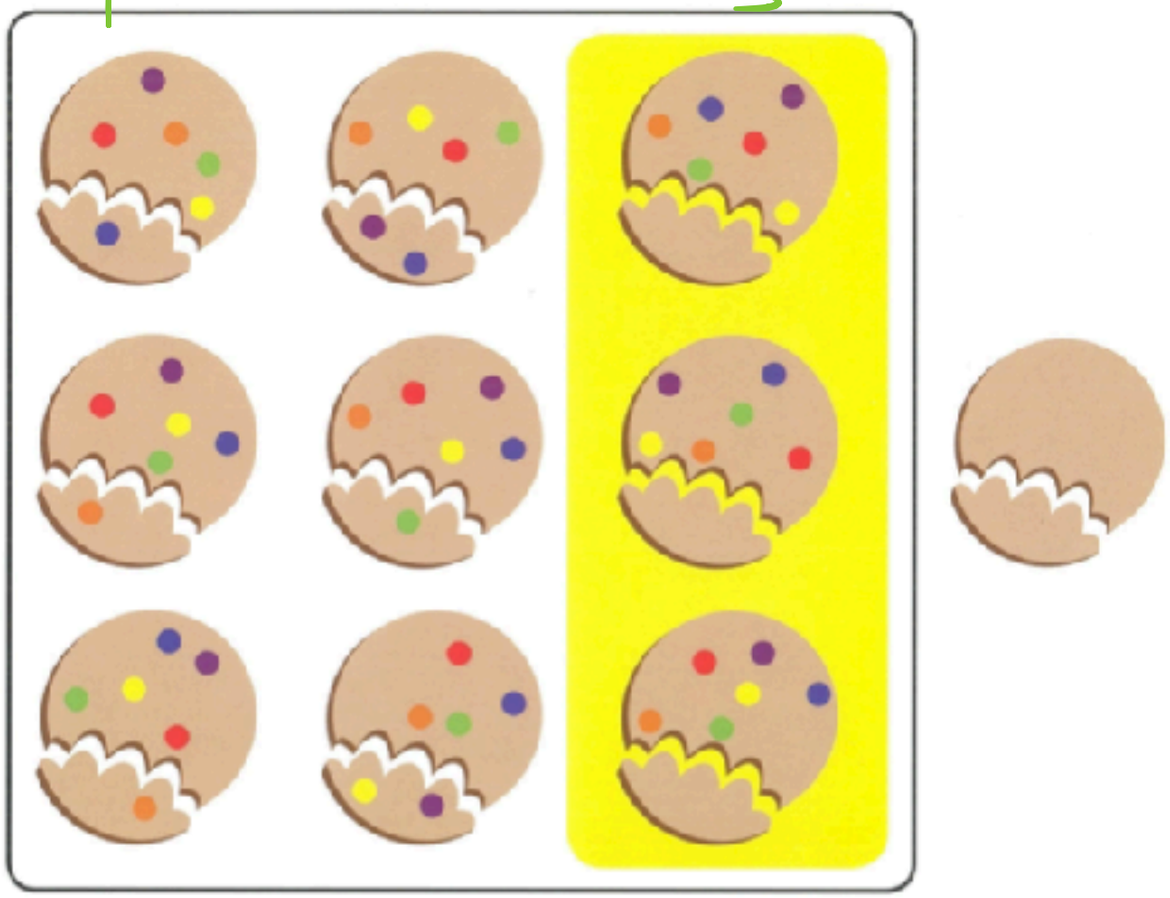
Take a bite of each.

There are 4 no-candy bites.
3 bites are from candy cookies.
1 bite is from a no-candy cookie.

$$p(B=0) = \frac{4}{10}$$

$$p(B=1) = \frac{6}{10}$$

$p(B=0 \mid C=1) = \frac{1}{3}$



Pr(|) = 3/4

1/3 of the candy cookies have a no-candy bite.

The probability of a candy cookie with a no-candy bite is 3/4.

Posterior prob

$$p(C=1 \mid B=0) = \frac{p(B=0 \mid C=1)\, p(C=1)}{p(B=0)}$$

$$= \frac{\frac{1}{3} \cdot \frac{9}{10}}{\frac{4}{10}} = \frac{3}{4}$$

This is the **prior distribution** of cookies.

$$p(C=1) = \frac{9}{10}$$

This is the **posterior distribution** of cookies.

$$p(C=1 \mid B=0) = \frac{3}{4}$$

Now you know
BAYESIAN PROBABILITY!

This bite probably came from a candy cookie!

Classification: pick the most probable option

University of Amsterdam
Informatics Institute

AMLAB Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 2.3 - Maximum Likelihood

Erik Bekkers

(Bishop 1.2.3 - 1.2.5)

Three Statistical Learning Principles:

Maximum Likelihood

Maximum A Posteriori

Bayesian Prediction

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Maximum Likelihood Estimation

- **Objective**: Find maximum likelihood solution $\mathbf{w}_{ML}$

$$\mathbf{w}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(D \,|\, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i}^{N} p(x_i \,|\, \mathbf{w})$$

- **Problem**: numerical underflow/overflow

Solution:

Maximum Likelihood Estimation

- **Objective**: Find maximum likelihood solution $\mathbf{w}_{ML}$

$$\mathbf{w}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(D \,|\, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i}^{N} p(x_i \,|\, \mathbf{w})$$

- **Problem**: numerical underflow/overflow

$\log(a \cdot b) = \log(a) + \log(b)$

- **Solution**: Maximize log-likelihood instead:

$$\mathbf{w}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i}^{N} p(x_i \,|\, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \log \prod_{i}^{N} p(x_i \,|\, \mathbf{w})$$

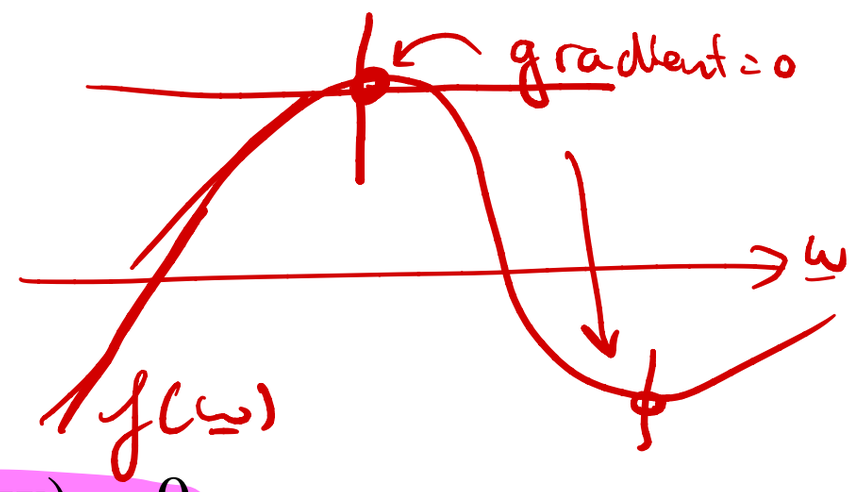
$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i \,|\, w)$$

- Or *minimize* error function negative log-likelihood function:

$$E(D \,|\, \mathbf{w}) = -\log p(D \,|\, \mathbf{w})$$

if we want to frame it as a minimization problem

Why log-trick?



- Solution to argmax problem is obtained by solving:

$$\underset{\mathbf{w}}{\operatorname{argmax}}\ p(D\,|\,\mathbf{w}) \quad \Leftrightarrow \quad \text{Solve: } \frac{\partial}{\partial \mathbf{w}} p(D\,|\,\mathbf{w}) = 0$$

- We can take the log as $p(x_i\,|\,\mathbf{w}) > 0$

$$\underset{\mathbf{w}}{\operatorname{argmax}}\ \log p(D\,|\,\mathbf{w}) \quad \Leftrightarrow \quad \text{Solve: } \frac{\partial}{\partial \mathbf{w}} \log p(D\,|\,\mathbf{w}) = 0$$

Same problem? Yes!

$\frac{\partial}{\partial w} \log p(D|w) = \frac{1}{p(D|w)} \cdot \frac{\partial}{\partial w} p(D|w) = 0$

recall $\frac{\partial}{\partial x} \log x = \frac{1}{x}$

Since $p(D|w)$ will never be zero. thus $\frac{\partial}{\partial w} p(D|w)$ has to be.

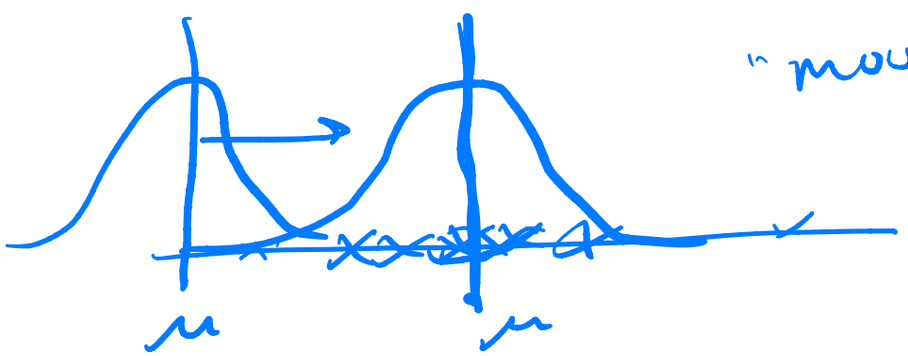
ML Estimator for Gaussian Distributions

- I.i.d. Gaussian distributed real variables $D = (x_1, x_2, \ldots, x_N)$

$$p(x \,|\, \mathbf{w}) = \mathcal{N}(x \,|\, \mu, \sigma^2) \quad \Rightarrow \quad p(D \,|\, \mathbf{w}) = (2\pi\sigma^2)^{-N/2} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

- Log likelihood

$$\log p(D \,|\, \mu, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{1}{2\sigma^2}(x_i - \mu)^2$$



- Estimate model parameters:

$$\mu_{ML}, \sigma_{ML} = \underset{\mu,\sigma}{\operatorname{argmax}} \log p(D \,|\, \mu, \sigma^2)$$

ML Estimator for Gaussian Distributions

- log likelihood:

$$\log p(D \mid \mu, \sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Maximum Likelihood solution for μ:

Steps: 1. compute derivative 2. set to zero 3. solve equation

$$\frac{\partial}{\partial \mu}\log p(D \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\frac{\partial}{\partial \mu}(x_i - \mu)^2 = 0$$

$$\Rightarrow -\frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^{N} x_i - \sum_{i=1}^{N}\mu = 0$$

$$\Rightarrow \sum_{i=1}^{N} x_i - N\mu = 0 \Rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

μ_{ML} is Sample mean:

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

ML Estimator for Gaussian Distributions

- log likelihood:

$$\log p(D\,|\,\mu,\sigma^2) = -\ \frac{N}{2}\log 2\pi\sigma^2\ -\ \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

- Maximum Likelihood solution for σ^2:

$$\frac{\partial}{\partial\sigma^2}\log p(D\,|\,\mu,\sigma^2) = -\frac{N}{2}\frac{1}{2\pi\sigma^2}\cdot 2\pi + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$\cdot 2\sigma^4 \quad \Rightarrow\ -N\sigma^2 + \sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$\Rightarrow\ \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

σ^2_{ML} is sample variance:

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

ML Estimator for Gaussian Distributions

- How well do the ML estimates represent the true parameters?

$$p(D\,|\,\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

- Drawing multiple datasets $D \sim p(D\,|\,\mu,\sigma^2)$, what is the *expected value* of μ_{ML}?

$D_1 = \{x_1^1, \ldots, x_N^1\}$, $D_2 = \{x_1^2, \ldots, x_N^2\}$

- ML estimate of the mean:

expectation over datasets

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\mu_{ML}] = \mathbb{E}_{D\sim p(D|\mu,\sigma^2)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right]$$

- Bias of estimator:

$$\mathbb{E}[\mu_{ML}] - \mu =$$

ML Estimator for Gaussian Distributions

- How well do the ML estimates represent the true parameters?

$$p(D\,|\,\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

- Drawing multiple datasets $D \sim p(D\,|\,\mu,\sigma^2)$, what is the *expected value* of μ_{ML}?

- ML estimate of the mean:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\mu_{ML}] = \mathbb{E}_{D\sim p(D|\mu,\sigma^2)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{x\sim p(x|\mu,\sigma^2)}\left[x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$$

- Bias of estimator:

$$\mathbb{E}[\mu_{ML}] - \mu = 0$$

ML Estimator for Gaussian Distributions

- How well do the ML estimates represent the true parameters?

$$p(D\,|\,\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

- Drawing multiple datasets $D \sim p(D\,|\,\mu,\sigma^2)$, what is the *expected value* of μ_{ML}?

- ML estimate of the mean:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\mu_{ML}] = \mathbb{E}_{D\sim p(D|\mu,\sigma^2)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{x\sim p(x|\mu,\sigma^2)}\left[x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$$

- Bias of estimator ML estimator for μ:

$$\mathbb{E}[\mu_{ML}] - \mu = 0$$

ML Estimator for Gaussian Distributions

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]$$

ML Estimator for Gaussian Distributions

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n)^2\right]$$

ML Estimator for Gaussian Distributions

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i^2 - \frac{2x_i}{N}\sum_{n=1}^{N}x_n + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}x_m x_n\right]$$

ML Estimator for Gaussian Distributions

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i^2 - \frac{2x_i}{N}\sum_{n=1}^{N}x_n + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}x_m x_n\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbb{E}\left[x_i^2\right] - \frac{2}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_i x_n\right] + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\mathbb{E}\left[x_m x_n\right]\right\}$$

$$= \ldots\ldots = \frac{N-1}{N}\sigma^2$$

$$\mathbb{E}[x_i x_j] = \begin{cases} \mu^2 + \sigma^2 & \text{if} \quad i = j \\ \mu^2 & \text{if} \quad i \neq j \end{cases}$$

← derive from known covariances!

Values for $\mathbb{E}[x_i x_j]$

- $i = j$:

$$\mathrm{Var}[x_i] = \mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2 = \sigma^2$$

$$\Rightarrow \quad \mathbb{E}[x_i x_j] = \mu^2 + \sigma^2$$

- $i \neq j$:

$$\mathrm{Cov}[x_i, x_j] = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j] = 0$$

$$\Rightarrow \quad \mathbb{E}[x_i x_j] = \mu^2$$

ML Estimator for Gaussian Distributions (VI)

- For data generated from

$$p(D\,|\,\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

- ML gives biased estimator for σ^2

$$\mathbb{E}[\sigma^2_{ML}] = \frac{N-1}{N}\sigma^2$$

apparently $\mathbb{E}[\sigma^2_{MC}] \neq \sigma^2$?

- Unbiased variance estimator:

$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{ML}$$

simply correct for it

Biased Maximum Likelihood Estimator

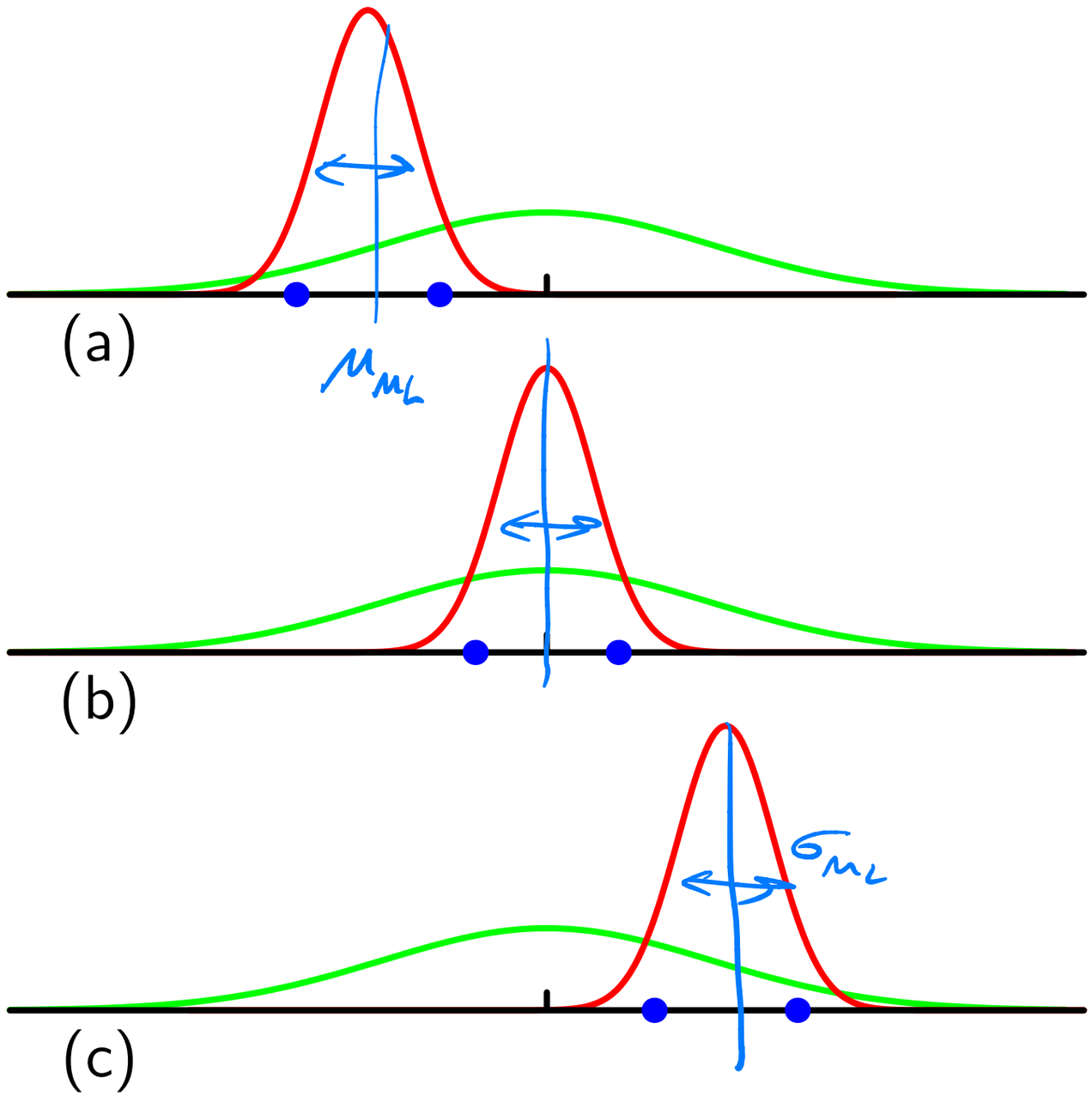


Figure: Bias in ML estimator for variance (Bishop 1.15)