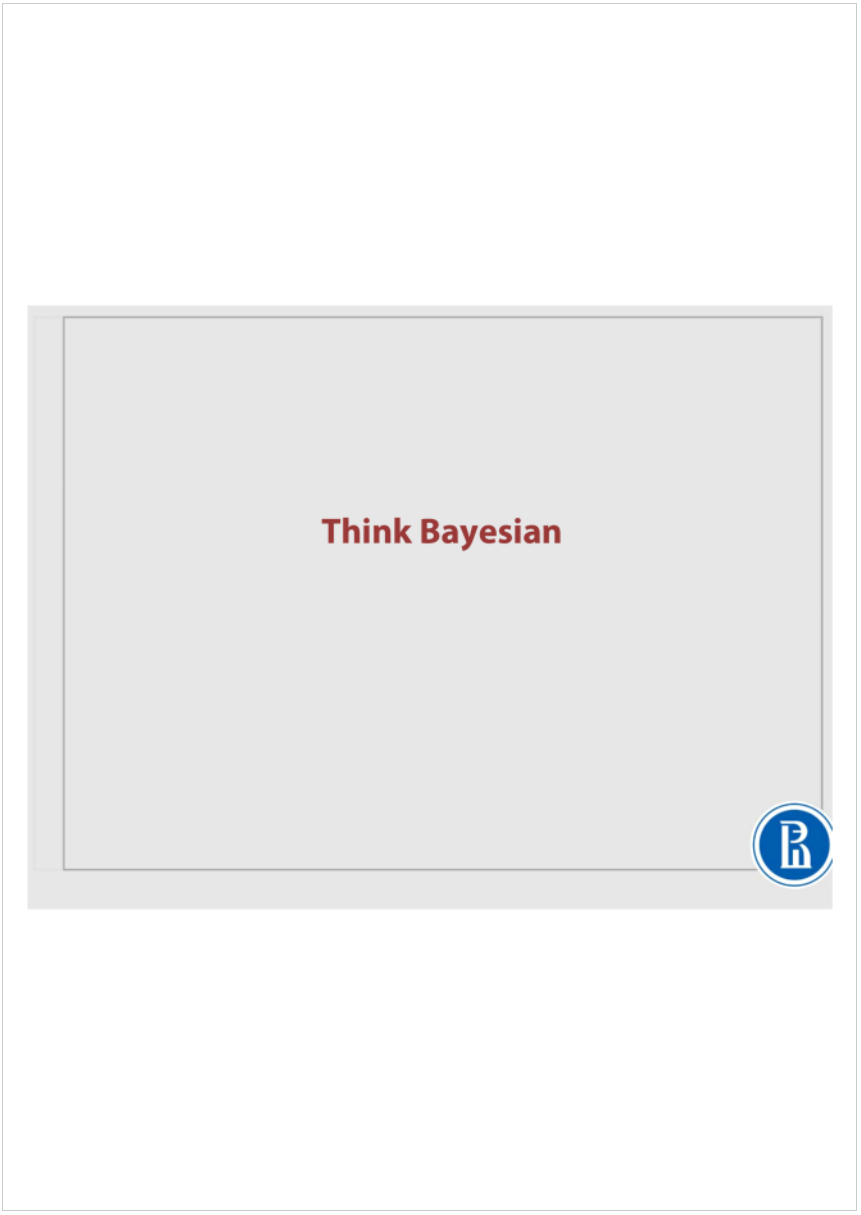
Intro to Bayesian

Friday, June 22, 2018 12:32 AM

Lecture slides



A man is running. Why?

Possible explanations:

1  He is in a hurry

2  He is doing sports

3  He always runs

4  He saw a dragon

A man is running. Why?

Principle 1: Use prior knowledge

1 He is in a hurry

2 He is doing sports

3 He always runs

4 He saw a dragon

A man is running. Why?

Principle 1: Use prior knowledge

1 
He is in a hurry

2 
He is doing sports

3 
He always runs

4 
~~He saw a dragon~~

Low prior probability

A man is running. Why?

Principle 2: Chose answer that explains observations the most

1 
He is in a hurry

2 
He is doing sports

3 
He always runs

4 
~~He saw a dragon~~

A man is running. Why?

Principle 2: Chose answer that explains observations the most

1 
He is in a hurry

2 
~~He is doing sports~~
Contradicts the data

3 
He always runs

4 
~~He saw a dragon~~

A man is running. Why?

Principle 3: Avoid making extra assumptions

1 
He is in a hurry

2 
~~He is doing sports~~

3 
He always runs

4 
~~He saw a dragon~~

A man is running. Why?

Principle 3: Avoid making extra assumptions

1 
He is in a hurry

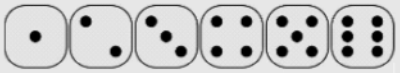
2 
~~He is doing sports~~

3 
~~He always runs~~

4 
~~He saw a dragon~~

Too many assumptions

Probability



$P(threw\ 5) = \frac{1}{6}$

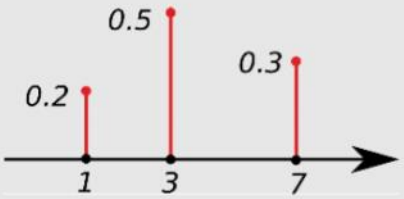
$P(threw\ odd) = \frac{1}{2}$

Random variables

Discrete

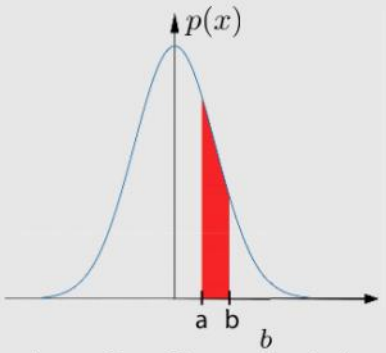
Continuous

Discrete: Probability Mass Function (PMF)



$$P(X) = \begin{cases} 0.2 & X = 1 \\ 0.5 & X = 3 \\ 0.3 & X = 7 \\ 0 & otherwise \end{cases}$$

Continuous: Probability Density Function (PDF)



$$P(x \in [a, b]) = \int_a^b p(x)dx$$

Independence

X and Y are independent if:

Marginals

$$P(X,Y) = P(X)P(Y)$$

Joint

Draws from the deck

52 cards

$$P(X_1 = 9\clubsuit, X_2 = 9\clubsuit) = 0$$

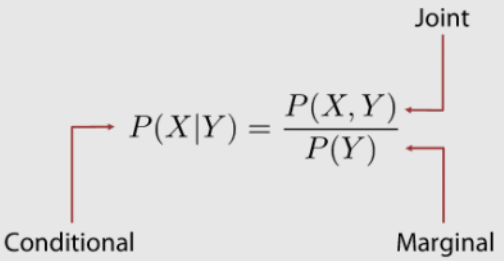
$$P(X_1 = 9\clubsuit)P(X_2 = 9\clubsuit) = \frac{1}{52^2}$$

Two coins

$$P(X_1 = \text{h}, X_2 = \text{t}) = P(X_1 = \text{h})P(X_2 = \text{t})$$

Conditional probability

Probability of **X** given that **Y** happened:



Conditional probability

$$P(M) = 0.4$$
$$P(M \,\&\, F) = 0.25$$

Midterm Final

$$P(F|M) = \frac{P(M \,\&\, F)}{P(M)} = \frac{0.25}{0.4} = 0.625$$

Chain rule

$$P(X, Y) = P(X|Y)P(Y)$$

Chain rule

$$P(X, Y) = P(X|Y)P(Y)$$

$$P(X, Y, Z) = P(X|Y, Z)P(Y|Z)P(Z)$$

Chain rule

$$P(X,Y) = P(X|Y)P(Y)$$

$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$

$$P(X_1,\ldots,X_N) = \prod_{i=1}^{N} P(X_i|X_1,\ldots,X_{i-1})$$

Sum rule

Marginalization

$$p(X) = \int_{-\infty}^{\infty} p(X, Y) dY$$

Bayes theorem

θ — parameters

X — observations

Posterior Likelihood Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)}$$

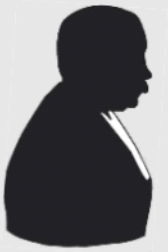
Evidence

Lecture slides (1)

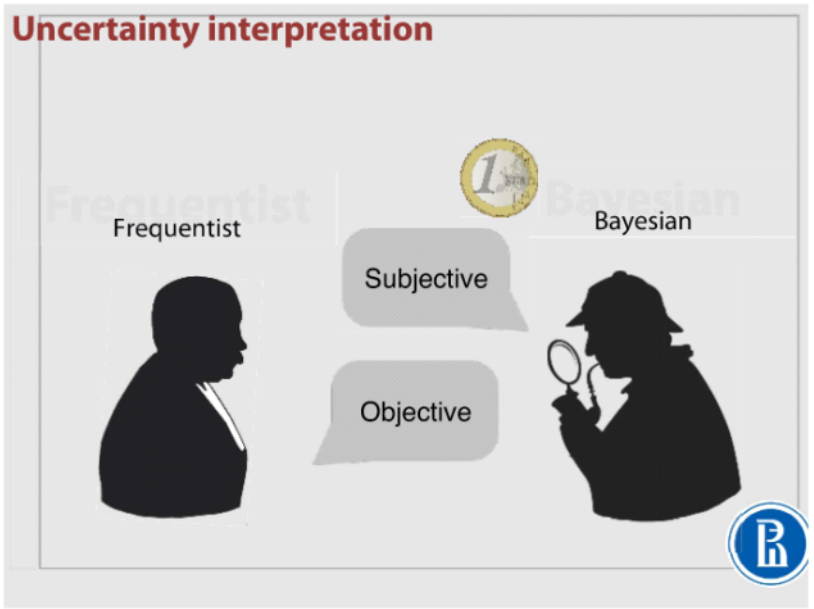
Bayesian approach to statistics

Different approaches to statistics

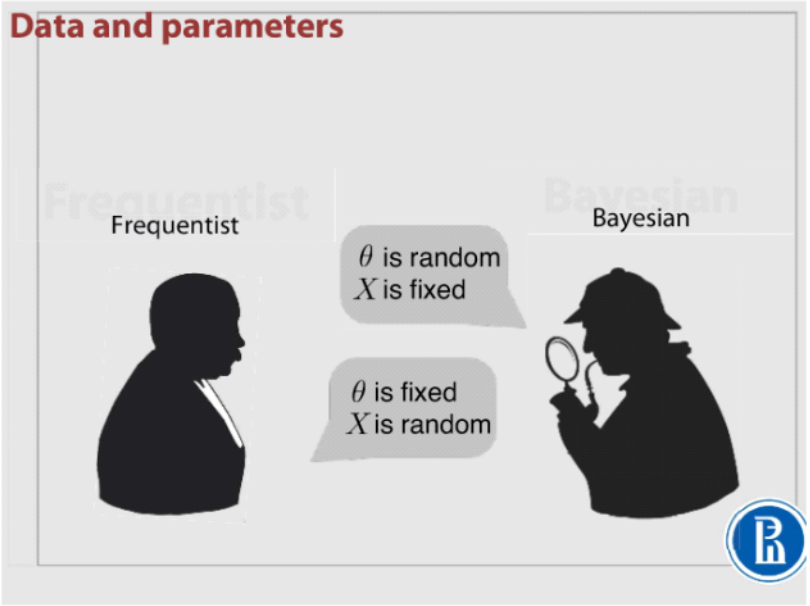
Frequentist

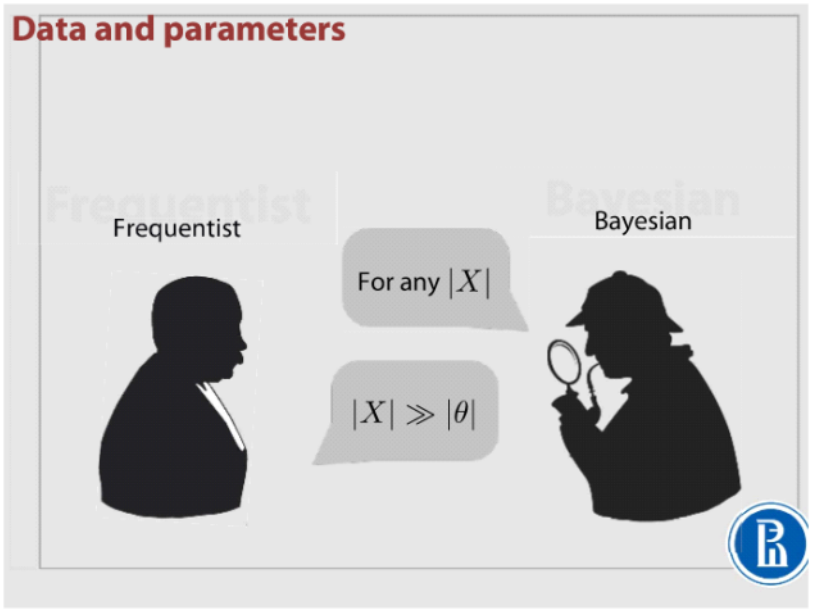


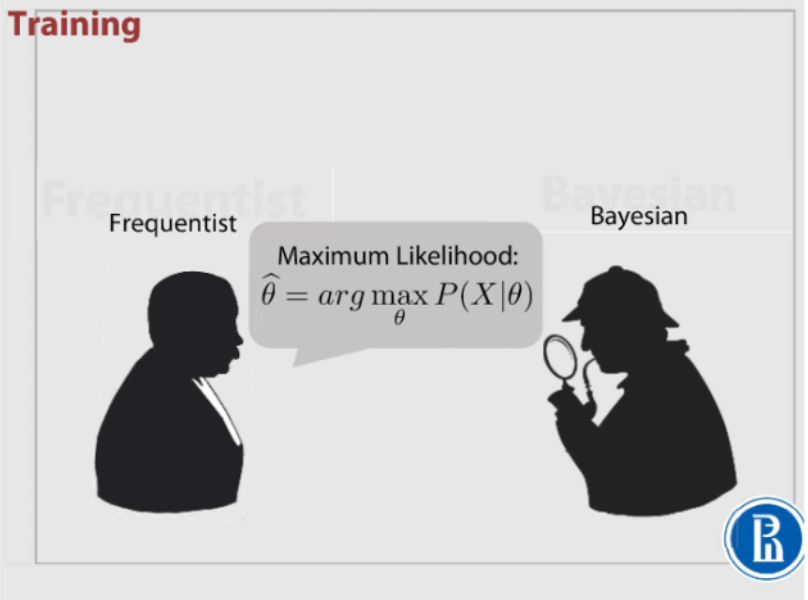
Bayesian

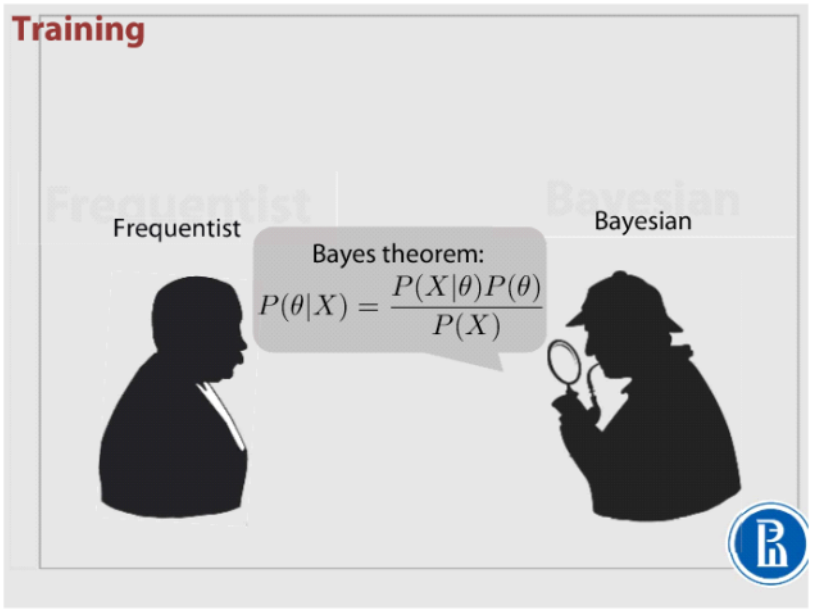

Uncertainty interpretation
Frequentist
Bayesian
Subjective
Objective

Data and parameters




Data and parameters
Frequentist
Bayesian
For any $|X|$
$|X| \gg |\theta|$


Training
Frequentist
Bayesian
Maximum Likelihood:
$\widehat{\theta} = arg\max_{\theta} P(X|\theta)$


Training
Frequentist
Bayesian
Bayes theorem:
$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$

Classification

Training:

$$P(\theta|X_{\text{tr}}, y_{\text{tr}}) = \frac{P(y_{\text{tr}}|X_{\text{tr}}, \theta)P(\theta)}{P(y_{\text{tr}}|X_{\text{tr}})}$$

Prediction:

$$P(y_{\text{ts}}|X_{\text{ts}}, X_{\text{tr}}, y_{\text{tr}}) = \int P(y_{\text{ts}}|X_{\text{ts}}, \theta)P(\theta|X_{\text{tr}}, y_{\text{tr}})d\theta$$

In bayesian approach, prediction is

- (•) a weighted average of output of our model for all possible values of parameters

Correct

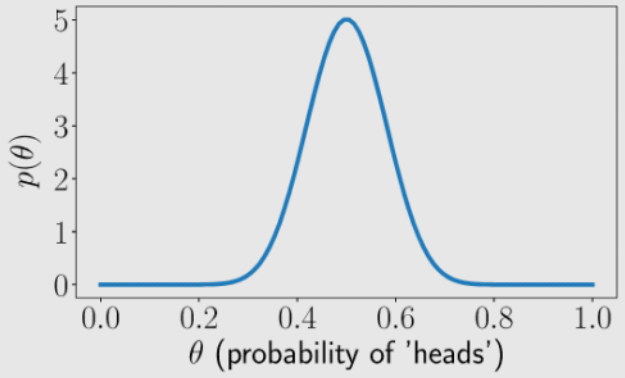
- ( ) a prediction of best-fitted values of parameters

Regularization

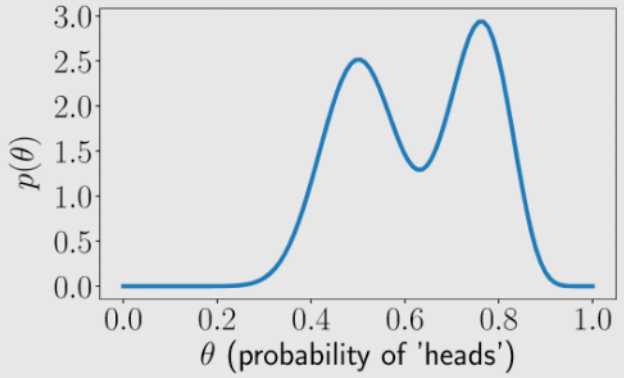
Regularizer

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Regularization



Regularization



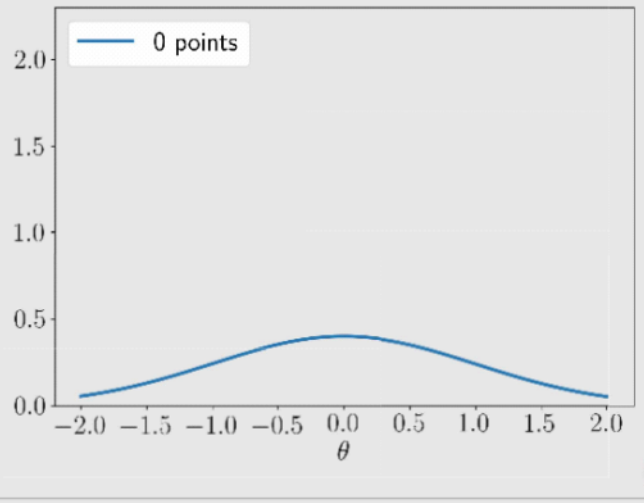
On-line learning

New prior — Likelihood — Prior

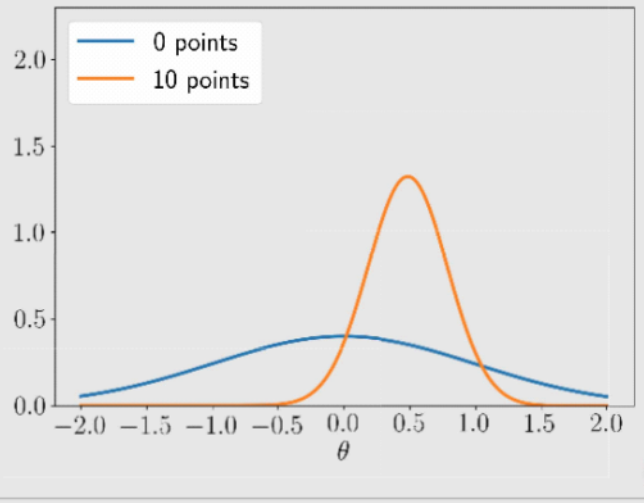
$$P_k(\theta) = P(\theta|x_k) = \frac{P(x_k|\theta)P_{k-1}(\theta)}{P(x_k)}$$

Posterior

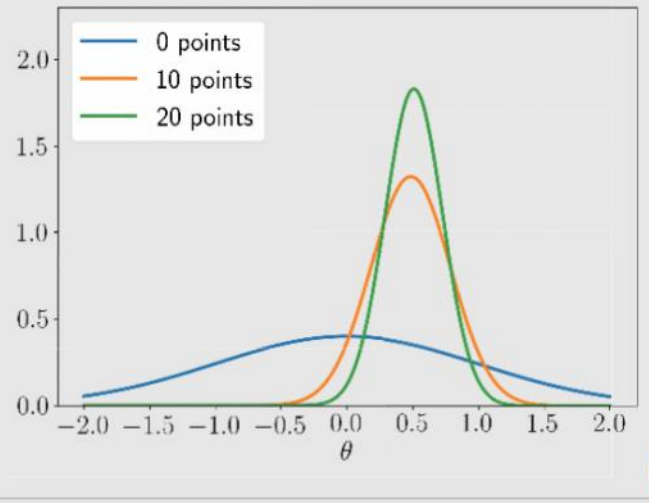
On-line learning

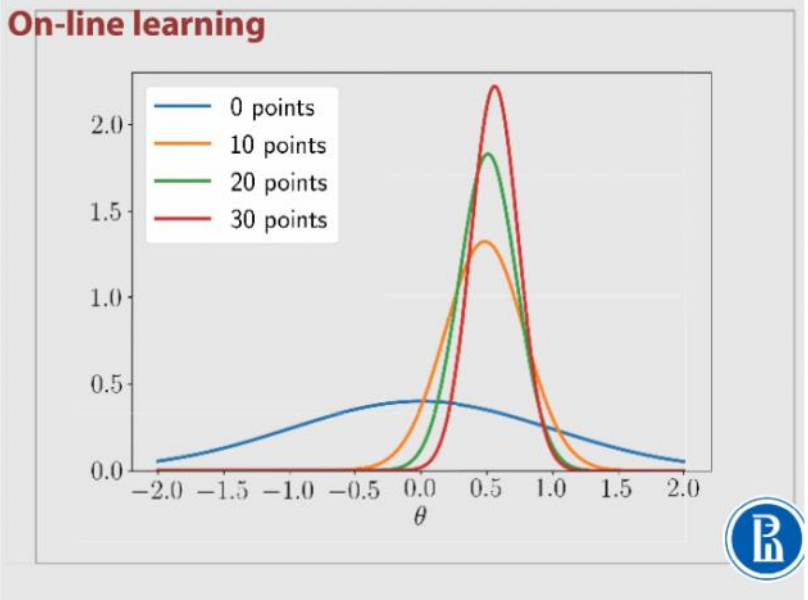


On-line learning

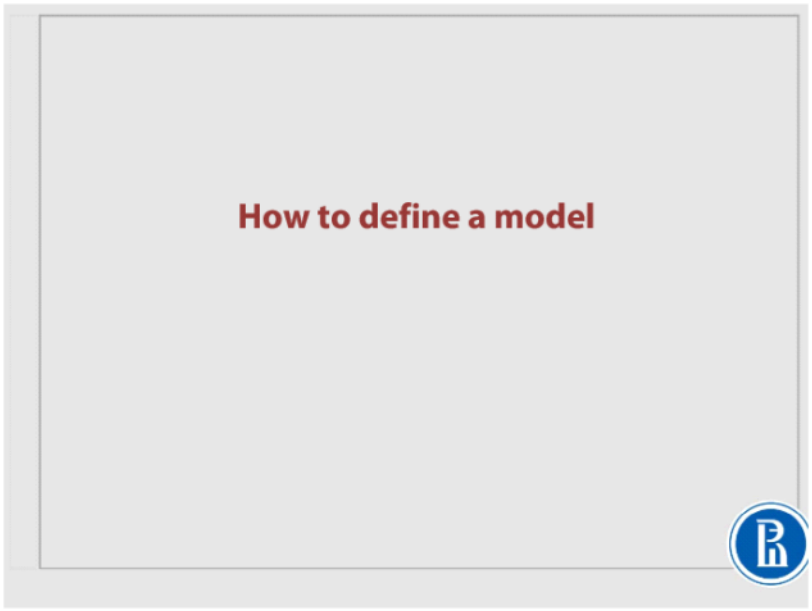


On-line learning





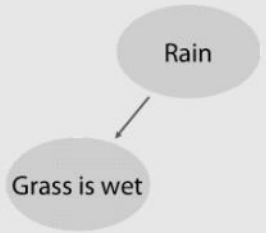
how to define a model


How to define a model

Bayesian network*

Nodes: random variables

Edges: direct impact

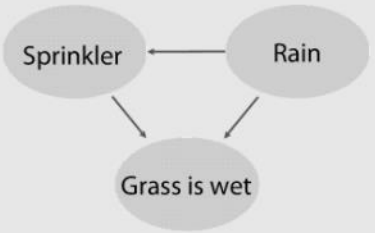


* Don't mix up with Bayesian neural network

Bayesian network*

Nodes: random variables

Edges: direct impact



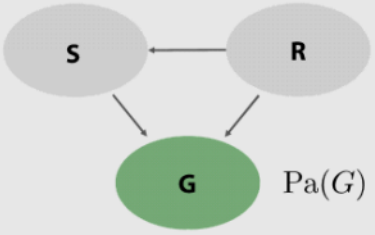
* Don't mix up with Bayesian neural network

Probabilistic model from BN

Model: joint probability over all variables

$$P(X_1, \ldots, X_n) = \prod_{k=1}^{n} P\Big(X_k | \mathrm{Pa}(X_k)\Big)$$

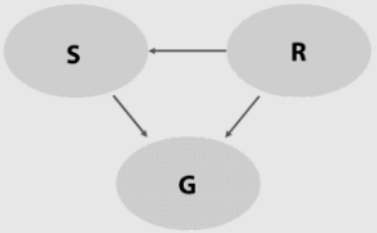
Parents



$\mathrm{Pa}(G) = \{R, S\}$

Probabilistic model from BN

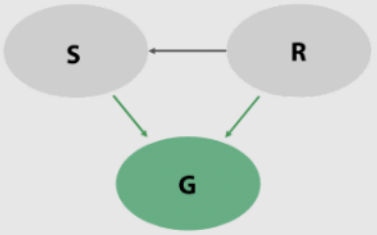
Model: joint probability over all variables



$$P(S, R, G) =$$

Probabilistic model from BN

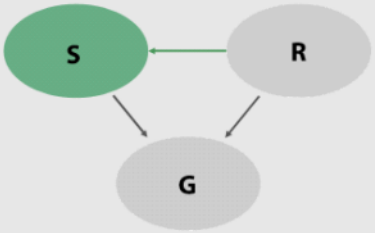
Model: joint probability over all variables



$$P(S, R, G) = P(G|S, R) \cdot$$

Probabilistic model from BN

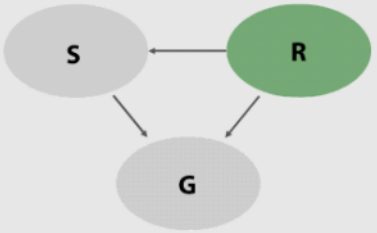
Model: joint probability over all variables



$$P(S, R, G) = P(G|S, R) \cdot P(S|R) \cdot$$

Probabilistic model from BN

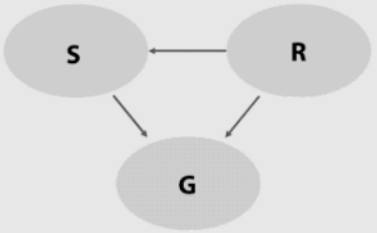
Model: joint probability over all variables



$$P(S, R, G) = P(G|S, R) \cdot P(S|R) \cdot P(R)$$

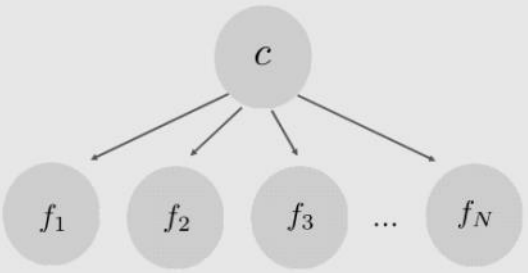
Probabilistic model from BN

Model: joint probability over all variables



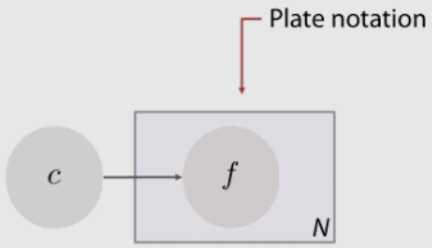
$$P(S, R, G) = P(G|S, R) \cdot P(S|R) \cdot P(R)$$

Naïve Bayes classifier



$$P(c, f_1, \ldots, f_N) = P(c) \prod_{i=1}^{N} P(f_i|c)$$

Naïve Bayes classifier



$$P(c, f_1, \ldots, f_N) = P(c) \prod_{i=1}^{N} P(f_i|c)$$

Can bayesian networks contain directed cycles?

- [x] No

Correct
We can't have interdependent variables. For those cases we can either join random variables into one random variable or use Markov Random Fields (MRF)

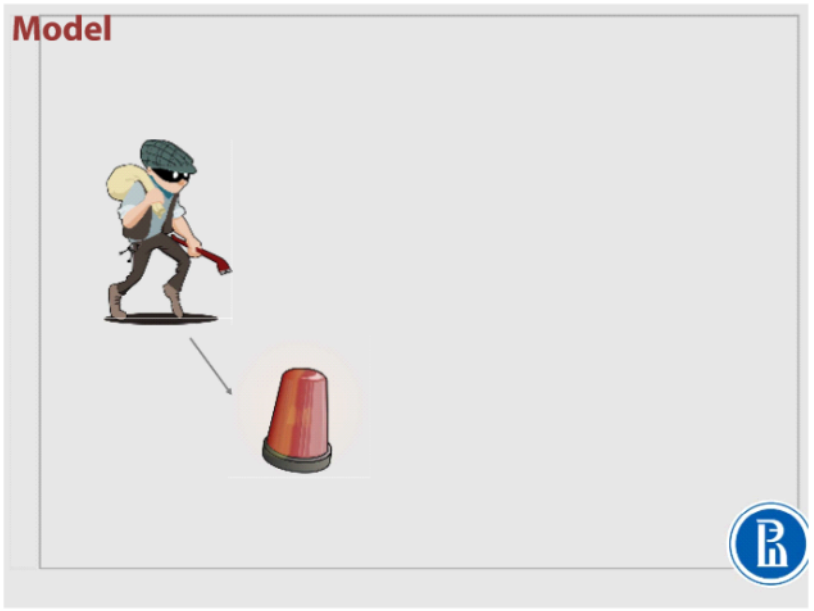
- [ ] Yes

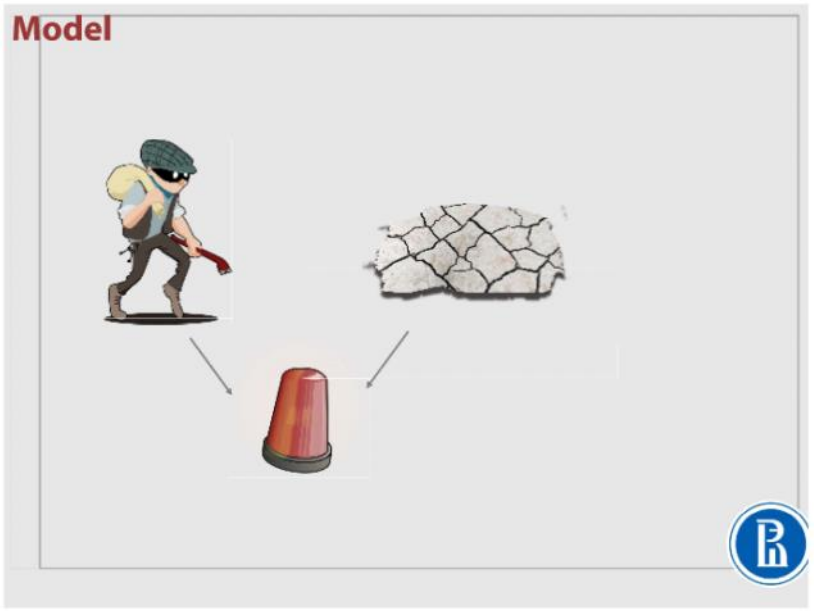
ex

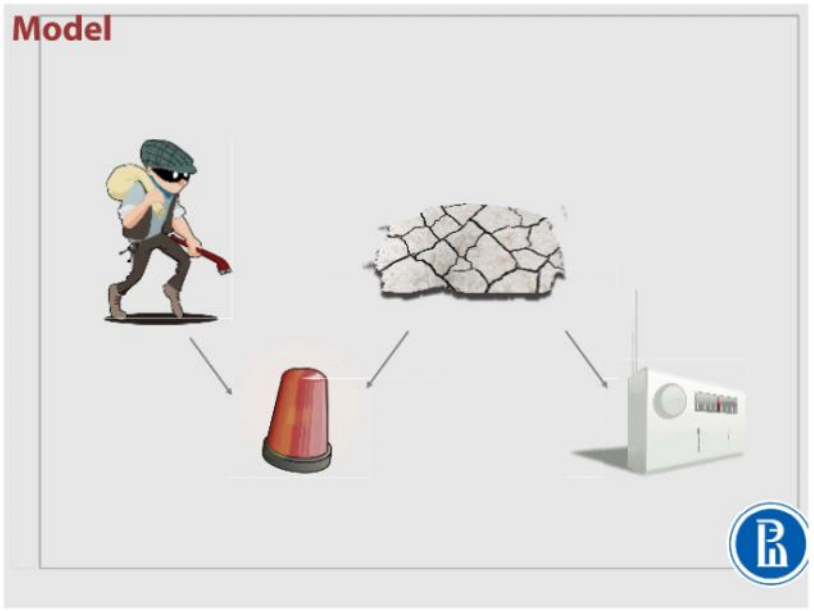
Example: thief & alarm


Model

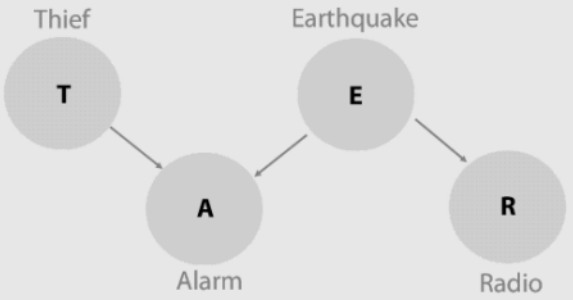

Model


Model


Model

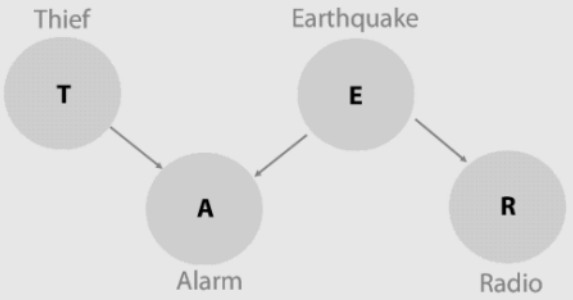

Model

Model



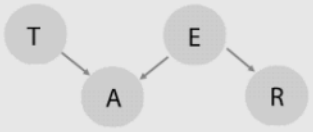
$$P(T, A, E, R) =$$

Model



$$P(T, A, E, R) = P(T)P(E)P(A|T, E)P(R|E)$$

Distributions




ТЕХНИЧЕСКИЙ СЛАЙД
Thief Earthquake
T
E
A
R

Distributions

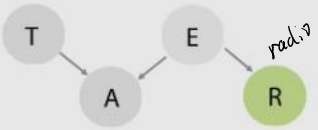
Priors	
$P(\mathbf{T} = \mathbf{1})$	$\mathbf{10^{-3}}$
$P(\mathbf{E} = \mathbf{1})$	$\mathbf{10^{-2}}$

Distributions

Priors	
$P(\mathbf{T}=\mathbf{1})$	$\mathbf{10^{-3}}$
$P(\mathbf{E}=\mathbf{1})$	$\mathbf{10^{-2}}$

$P(\mathbf{A}=\mathbf{1}\|\mathbf{T},\mathbf{E})$	$\mathbf{E}=\mathbf{0}$	$\mathbf{E}=\mathbf{1}$
$\mathbf{T}=\mathbf{0}$	0	1/10
$\mathbf{T}=\mathbf{1}$	1	1

Distributions

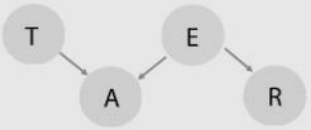


Priors	
$P(\mathbf{T}=\mathbf{1})$	$\mathbf{10^{-3}}$
$P(\mathbf{E}=\mathbf{1})$	$\mathbf{10^{-2}}$

$P(\mathbf{A}=\mathbf{1}\|\mathbf{T},\mathbf{E})$	$\mathbf{E}=\mathbf{0}$	$\mathbf{E}=\mathbf{1}$
$\mathbf{T}=\mathbf{0}$	0	1/10
$\mathbf{T}=\mathbf{1}$	1	1

$P(\mathbf{R}\|\mathbf{E})$	
$\mathbf{E}=\mathbf{0}$	0
$\mathbf{E}=\mathbf{1}$	1/2

Distributions



Priors	
$P(\mathbf{T}=\mathbf{1})$	$\mathbf{10^{-3}}$
$P(\mathbf{E}=\mathbf{1})$	$\mathbf{10^{-2}}$

$P(\mathbf{A}=\mathbf{1}\|\mathbf{T},\mathbf{E})$	$\mathbf{E}=\mathbf{0}$	$\mathbf{E}=\mathbf{1}$
$\mathbf{T}=\mathbf{0}$	0	1/10
$\mathbf{T}=\mathbf{1}$	1	1

$P(\mathbf{R}\|\mathbf{E})$	
$\mathbf{E}=\mathbf{0}$	0
$\mathbf{E}=\mathbf{1}$	1/2

$P(A|T,E)P(T)P(E) = 10^{-2} \times 10^{-3} = 10^{-5}$

$$P(T|A) = \frac{P(T,A)}{P(A)} = \frac{P(T,A,E) + P(T,A,\bar{E})}{P(T,A,E) + P(T,A,\bar{E}) + P(\bar{T},A,E) + P(\bar{T},A,\bar{E})} \approx 0.50$$

$P(A|\bar{T},\bar{E})P(\bar{T})P(\bar{E}) \simeq 0$

<< Part on marker board >> (8 mins)

$$P(T,A,\bar{E}) = P(A|T,\bar{E})P(T)P(\bar{E})$$
$$= 1 \times 10^{-3} \times (1-10^{-3})$$
$$= 0.00099$$

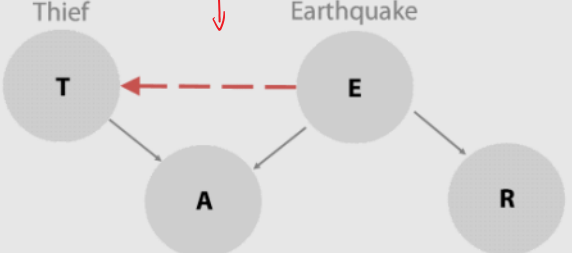
$$P(T|A,R) = \frac{P(T,A,R)}{P(A,R)}$$
$$= \frac{P(T,A,R,E) + P(T,A,R,\bar{E})}{P(A,R,T,E) + P(A,R,T,\bar{E}) + P(A,R,\bar{T},R) + P(A,R,\bar{T},\bar{R})}$$

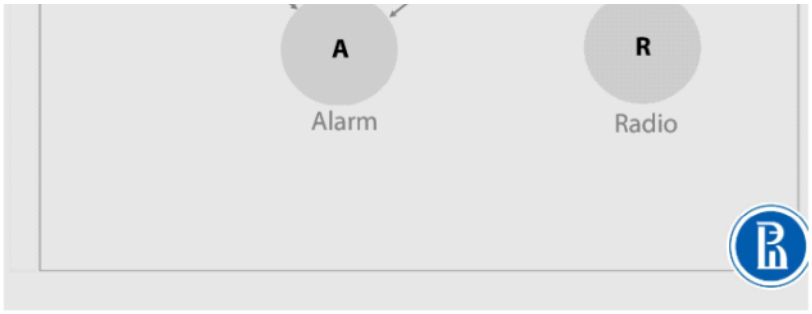
$P(A|T,E) \times P(R|E) \times P(T) \times P(E)$

$$P(T,A,R,\bar{E}) = P(A|T,\bar{E}) \times P(R|\bar{E}) \times P(T) \times P(\bar{E})$$
$$= 1 \times 0 \cdots$$
$$= 0$$

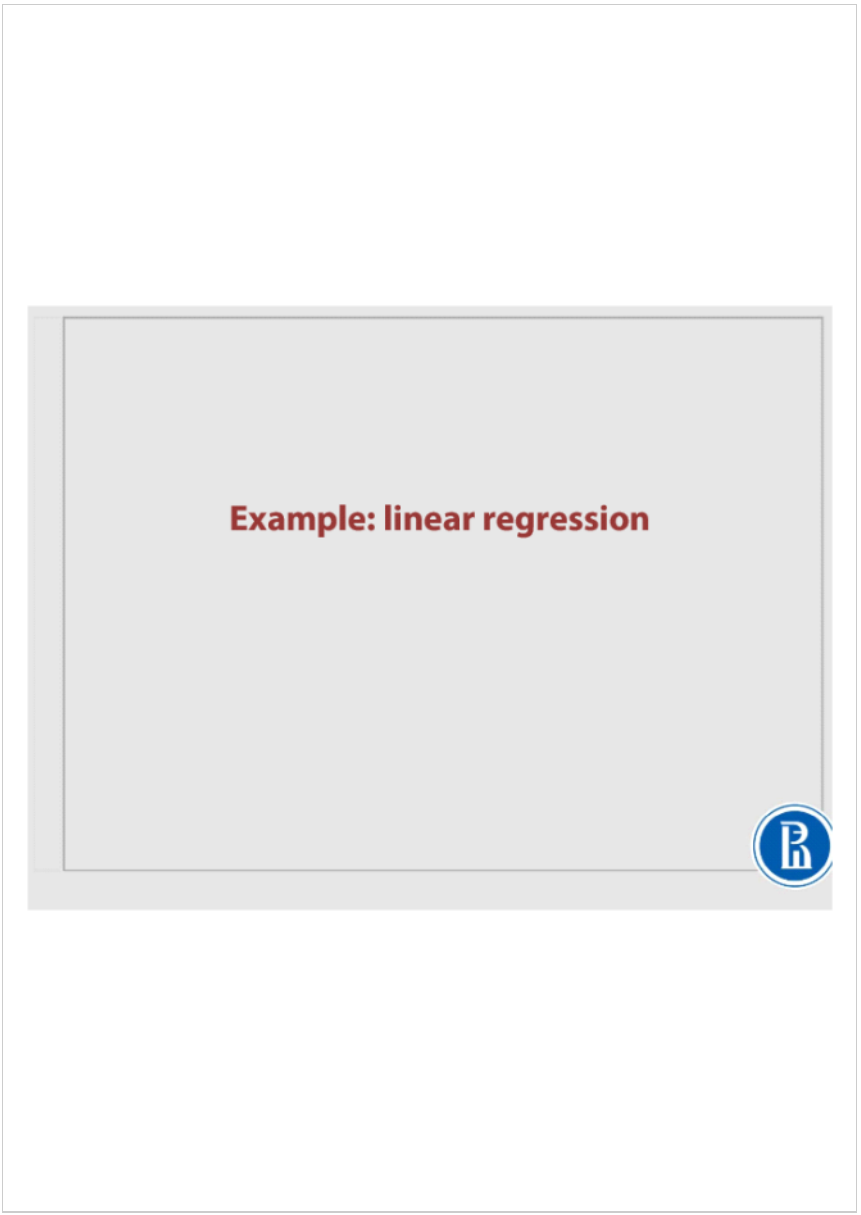
Correct model

revision of model. thief and earthquake are related.

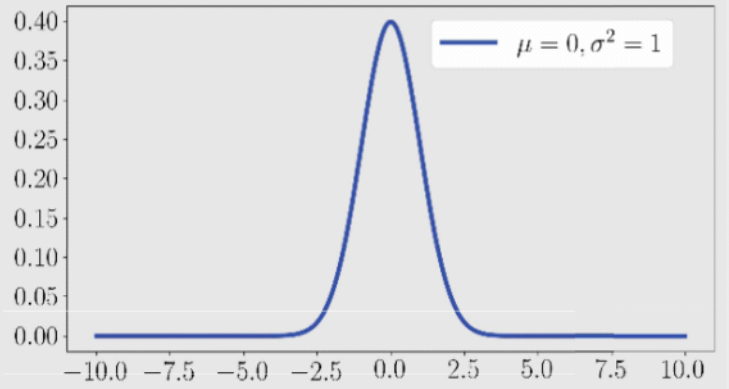




Im

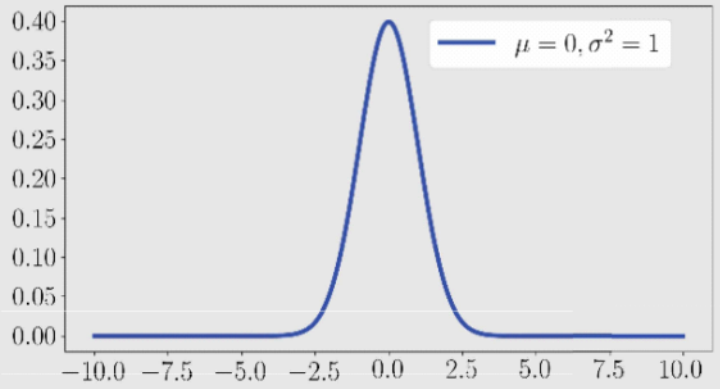

Example: linear regression

Univariate normal



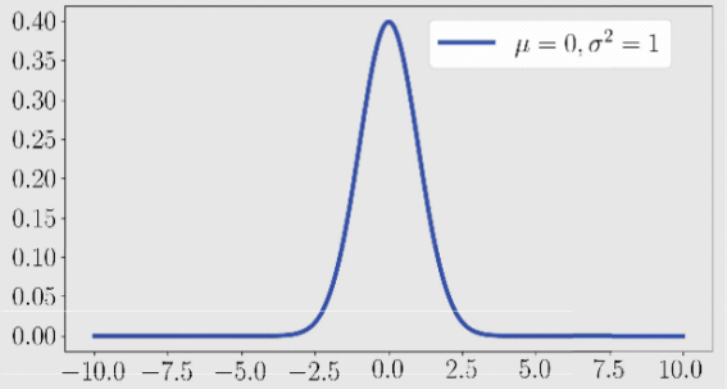
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Univariate normal



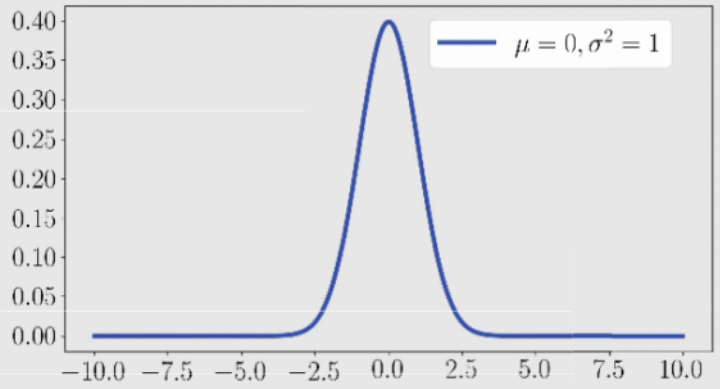
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Univariate normal



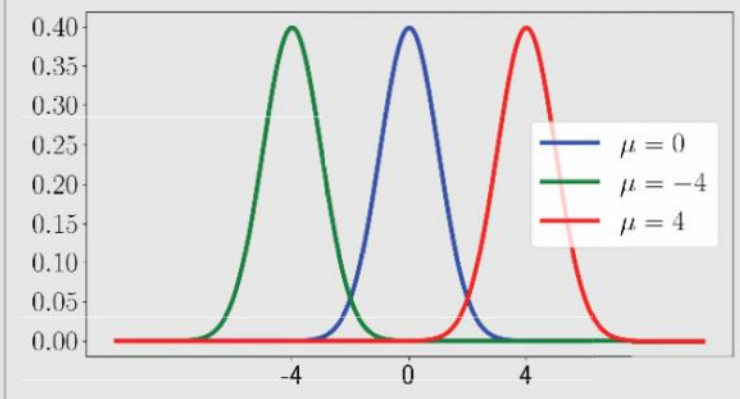
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Univariate normal



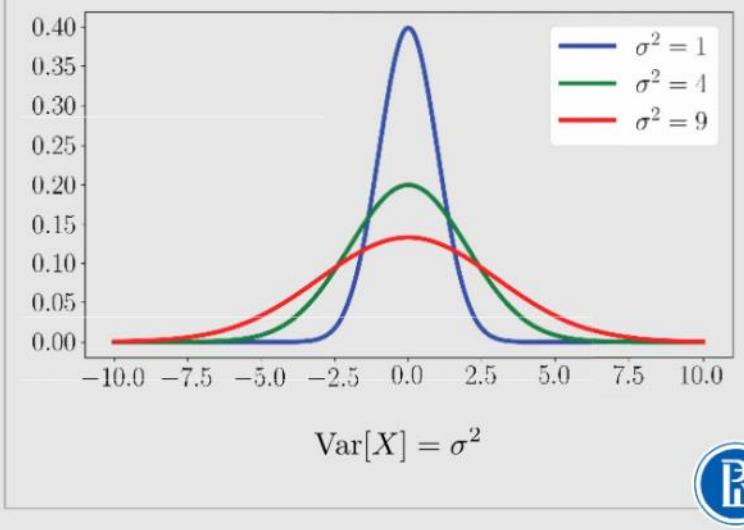
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Univariate normal: mean



$$\mathbb{E}X = \mu$$

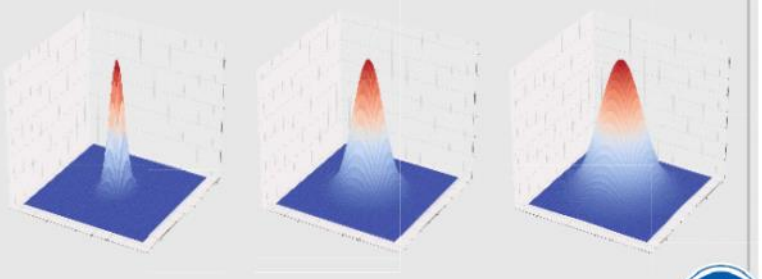
Univariate normal: variance



$$\mathrm{Var}[X] = \sigma^2$$

Multivariate normal

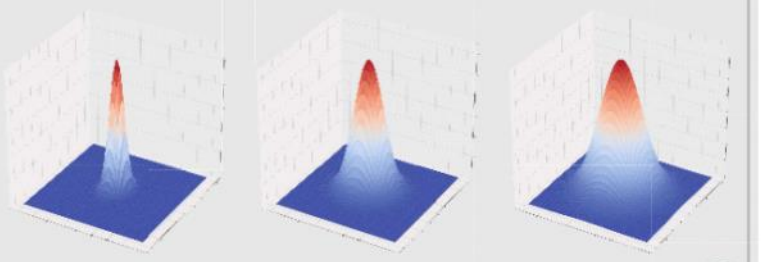
$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$



Multivariate normal

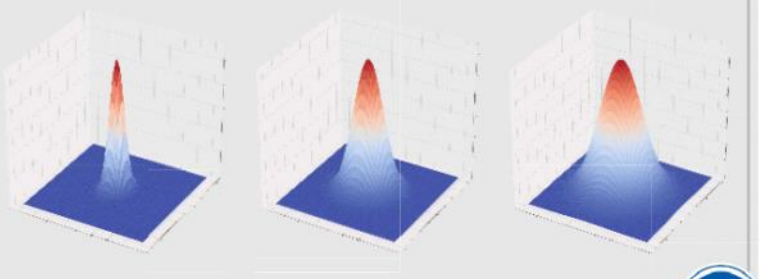
$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

$$\mathbb{E}X = \mu \quad \mathrm{Cov}[X] = \Sigma$$



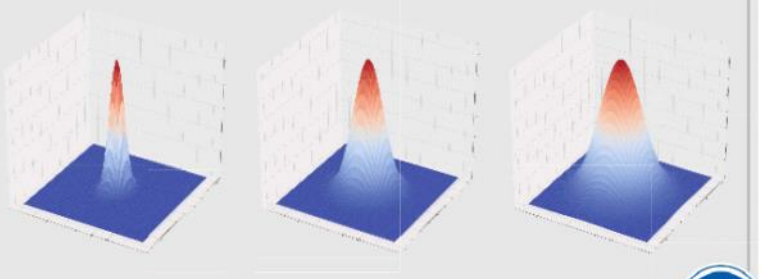
Multivariate normal

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



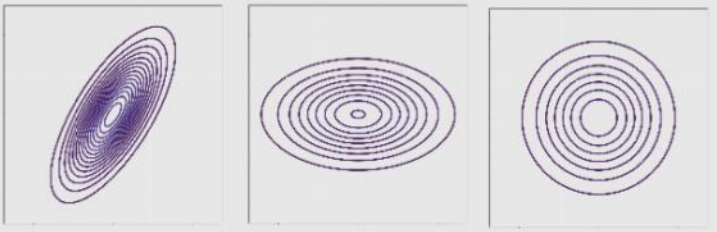
Multivariate normal

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



Multivariate normal

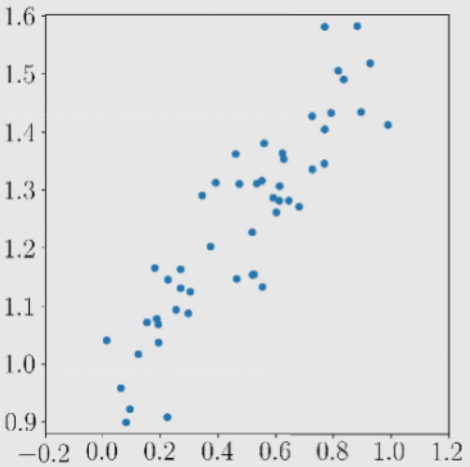
$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$



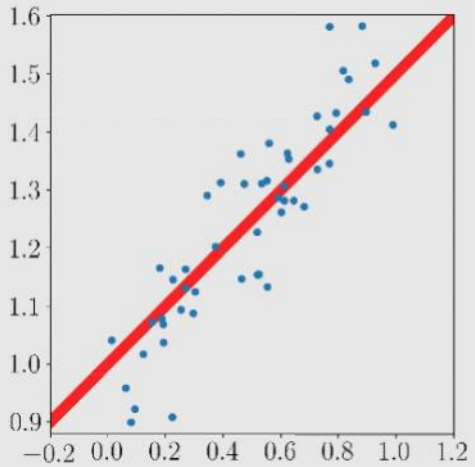
Full | Diagonal | Spherical

Parameters: $\frac{D(D+1)}{2}$ | Parameters: D | Parameters: 1

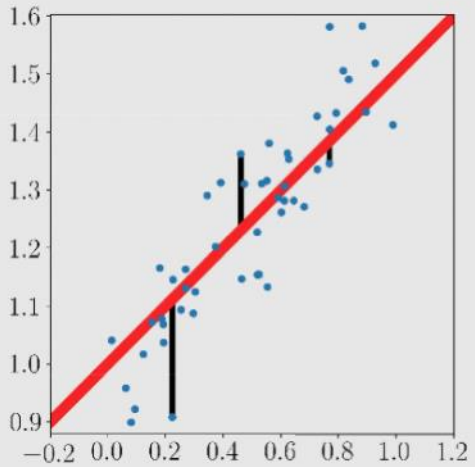
Linear regression



Linear regression

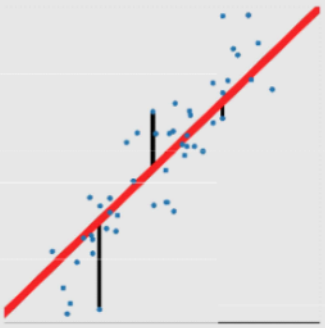


Linear regression



Least squares problem

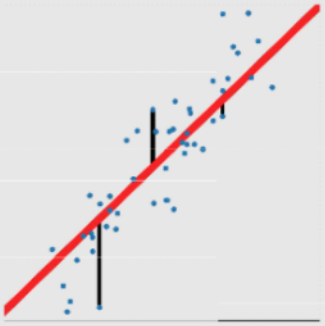
$$L(w) = \sum_{i=1}^{N} (w^T x_i - y_i)^2 = \left\| w^T X - y \right\|^2 \to \min_w$$



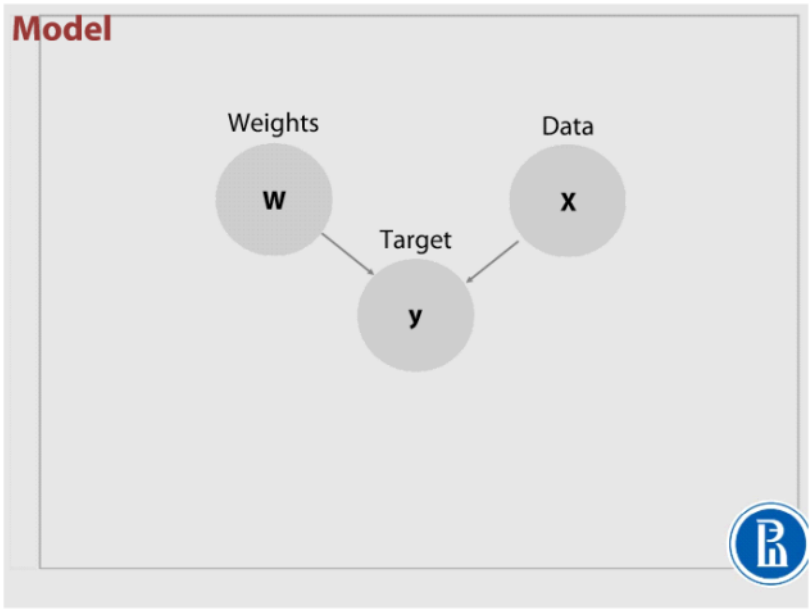
Least squares problem

total sum of squares

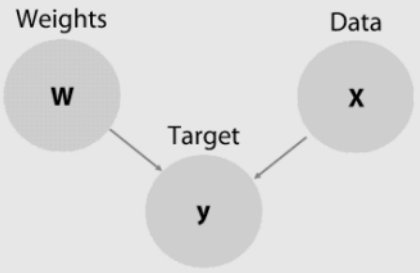
$$L(w) = \sum_{i=1}^{N} (w^T x_i - y_i)^2 = \left\| w^T X - y \right\|^2 \to \min_w$$



$$\widehat{w} = arg\min_w L(w)$$


Model
Weights
Data
w
x
Target
y

Model



$$P(w, y|X) = P(y|X, w)P(w)$$

$P(w,y|X) = P(y|X,w)P(w)$

$P(y|w,X) = \mathcal{N}(y|w^T X, \sigma^2 I)$.

$P(w) = \mathcal{N}(w, 0, \tau^2 I)$.

$P(w|y,X) \propto \frac{P(y,w|X)}{P(y|X)} \to \max_w$.

$\log[P(y,w|X)] = \log[P(y|X,w)P(w)] \to \max_w$.

$\log P(y|X,w) + \log P(w)$

$= \log c_1 \exp\left(-\frac{1}{2}(y - w^T X)^T (\sigma^2 I)(y - w^T X)\right)$

$+ \log c_2 \exp\left(-\frac{1}{2} w^T (\tau^2 I) w\right)$

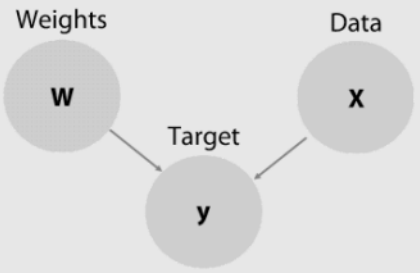
$\mathcal{L} = \frac{1}{2\sigma^2}(y - w^T X)^T (y - w^T X) - \frac{1}{2\tau^2} w^T w \Rightarrow \max_w$

$\min \|y - w^T X\|^2 + \lambda \|w\|^2$ ⇐ $\frac{1}{2}$ minimization problem.

least square — L2 regularization

Turn from least square to L2 regularized linear regression

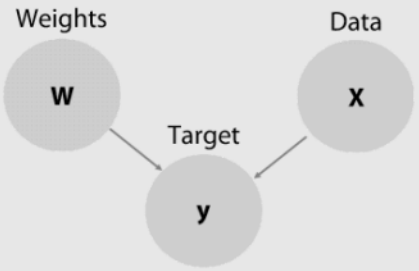
Model



$P(w, y|X) = P(y|X, w)P(w)$

$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$

Model



$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$
$$P(w) = \mathcal{N}(w|0, \gamma^2 I)$$

Please wait while OneNote loads this Printout...

Training ТЕХНИЧЕСКИЙ СЛАЙД (НА ДОСКЕ)

$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$
$$P(w) = \mathcal{N}(w|0, \gamma^2 I)$$

$$P(w|y, X) = \frac{P(w, y|X)}{P(y|X)} \propto P(w, y|X)$$

Training ТЕХНИЧЕСКИЙ СЛАЙД (НА ДОСКЕ)

$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$
$$P(w) = \mathcal{N}(w|0, \gamma^2 I)$$

$$P(w|y, X) = \frac{P(w, y|X)}{P(y|X)} \propto P(w, y|X)$$

$$P(w, y|X) \to \max_w$$

Training ТЕХНИЧЕСКИЙ СЛАЙД (НА ДОСКЕ)

$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$
$$P(w) = \mathcal{N}(w|0, \gamma^2 I)$$

$$P(w|y, X) = \frac{P(w, y|X)}{P(y|X)} \propto P(w, y|X)$$

$$P(w, y|X) \to \max_w \Leftrightarrow \log P(w, y|X) \to \max_w$$

Training ТЕХНИЧЕСКИЙ СЛАЙД (НА ДОСКЕ)

$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$
$$P(w) = \mathcal{N}(w|0, \gamma^2 I)$$

$$\log P(w, y|X) \to \max_w$$

Training ТЕХНИЧЕСКИЙ СЛАЙД (НА ДОСКЕ)

$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$
$$P(w) = \mathcal{N}(w|0, \gamma^2 I)$$

$$\log P(w, y|X) \to \max_w$$
$$-\frac{1}{2\sigma^2}||w^T X - y||^2 - \frac{1}{2\gamma^2}||w||^2 \to \max_w$$

Training ТЕХНИЧЕСКИЙ СЛАЙД (НА ДОСКЕ)

$$P(w, y|X) = P(y|X, w)P(w)$$
$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 I)$$
$$P(w) = \mathcal{N}(w|0, \gamma^2 I)$$

$$\log P(w, y|X) \to \max_w$$
$$-\frac{1}{2\sigma^2}||w^T X - y||^2 - \frac{1}{2\gamma^2}||w||^2 \to \max_w$$
$$||w^T X - y||^2 + C||w||^2 \to \min_w$$

mle

Maximum Likelihood Estimate

Consider a model parametrized by a vector θ, and let $X = (x_1, \dots x_N)$ be observed data samples from the model. Then the function $p(X|\theta)$ is called the *likelihood function* if viewed as a function of the parameter vector θ. It shows how probable the observed data X is for different values of θ. Note that the likelihood is not a probability distribution over θ (its integral with respect to θ may not be equal to one).

For example, if we consider the set X of independently drawn samples from the normal distribution with unknown parameters, then $\theta = (\mu, \sigma^2)$, and the likelihood function is

$$p(X|\theta) = \mathcal{N}(X|\mu, \sigma) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right)$$

treated as a function of μ and σ.

The Maximum Likelihood Estimate (MLE) for parameter is the value of θ which maximizes the likelihood. It is a very common way in statistics to estimate the unknown parameters for the model after observing the data.

Continuing the example above, let us find the MLE for parameter μ. As we assumed that samples are drawn independently from the model, the likelihood takes the form of a *product* of individual likelihood functions for each sample $p(x_i|\theta)$. When finding the MLE it is often convenient to find the maximum of the function $\log p(X|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$ (which in its turn, takes the form of the *sum* of individual log likelihood functions) instead of directly optimizing $p(X|\theta)$:

$$\log p(X|\theta) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

To maximize this expression with respect to μ, we set the partial derivative with respect to μ to zero and obtain:

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

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

Let us also consider multidimensional case for this problem: now each x_i is a d-dimensional vector drawn from the multivariate normal distribution with parameters mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Similarly, the log likelihood for this model takes the form:

$$\log p(X|\mu, \Sigma) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log\det\Sigma - \sum_{i=1}^{N}\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) =$$

$$= -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T\Sigma^{-1}x_i - \mu^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu\right) =$$

[use the fact that Σ is symmetric, thus Σ^{-1} is also symmetric which leads to:

$$\mu^T\Sigma^{-1}x_i = \left(\mu^T\Sigma^{-1}x_i\right)^T = x_i^T\left(\Sigma^{-1}\right)^T\mu = x_i^T\Sigma^{-1}\mu\Big]$$

$$= -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu\right)$$

Now to obtain the MLE for μ, we need to compute the derivative of this expression with respect to vector μ and set it to zero. We will use the following vector differentiation rules:

$$\frac{\partial}{\partial y}(a^T y) = a \quad \text{for } y \in \mathbb{R}^d, a \in \mathbb{R}^d$$

$$\frac{\partial}{\partial y}(y^T A y) = 2Ay \quad \text{for } y \in \mathbb{R}^d \text{ and symmetric matrix } A \in \mathbb{R}^{d \times d}$$

Applying them to the log likelihood expression, we get:

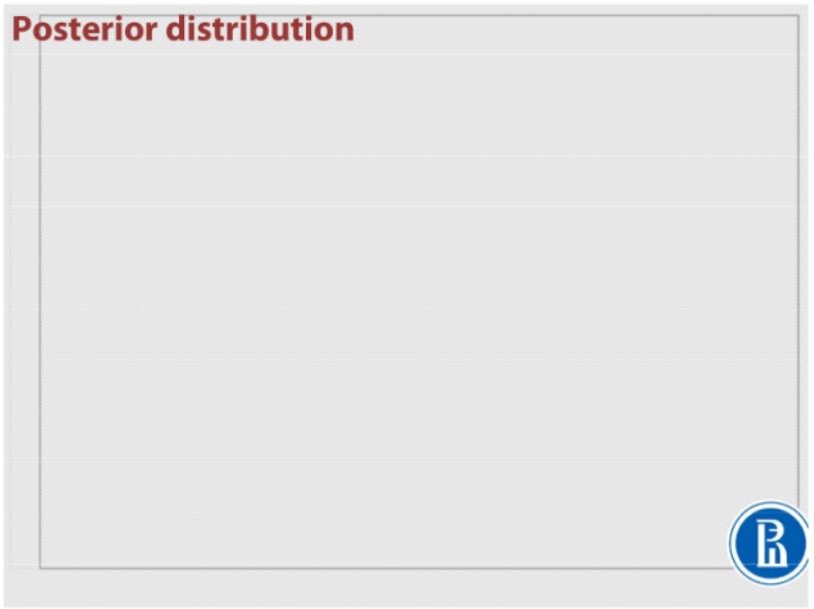
$$\frac{\partial}{\partial \mu}\log p(X|\mu, \Sigma) = -\frac{1}{2}\sum_{i=1}^{N}\left(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu\right) = \sum_{i=1}^{N}\Sigma^{-1}(x_i - \mu) = 0$$

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

analytical inference

Analytical inference

Posterior distribution



Posterior distribution

Likelihood — Prior

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Evidence

Posterior distribution

Likelihood — Prior

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Evidence

What is $P(X)$?

Posterior distribution

Likelihood — Prior

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Evidence

What is $P(X)$?

Van Gogh Starry night

Posterior distribution

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Likelihood, Prior, Evidence

What is $P(X)$?

Van Gogh Starry night

Van Gogh, Starry night over the Rhone

Imagine that you are training a neural network to play games. X is an image of the game screen and θ are network parameters. What statements are correct?

- [x] If we knew $P(X)$, we could generate new game-like frames

  Correct

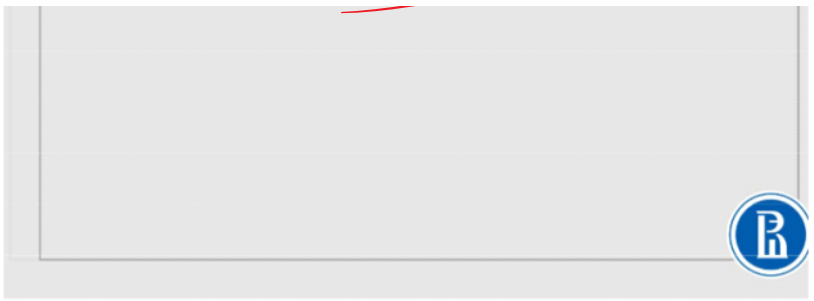
- [ ] $P(X)$ is a uni-modal Gaussian distribution

  Un-selected is correct

- [x] For some games, like "snake" modeling $P(X)$ may not be hard

  Correct

Maximum a posteriori



Maximum a posteriori

Maximum a posteriori

$$\theta_{\text{MP}} = arg \max_{\theta} P(\theta|X)$$

Maximum a posteriori

$$\theta_{\text{MP}} = arg\max_{\theta} P(\theta|X)$$

$$\theta_{\text{MP}} = arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)}$$

Maximum a posteriori

$$\theta_{\text{MP}} = arg\max_{\theta} P(\theta|X)$$

$$\theta_{\text{MP}} = arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\theta_{\text{MP}} = arg\max_{\theta} P(X|\theta)P(\theta)$$

为什么把 normalizing constant 去掉了？

Maximum a posteriori

$$\theta_{\text{MP}} = arg\max_{\theta} P(\theta|X)$$

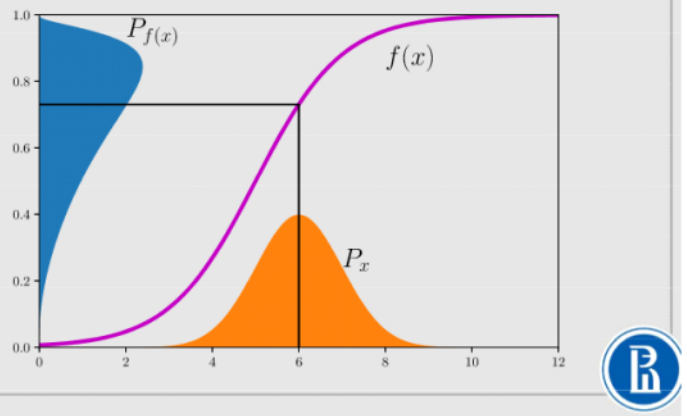
$$\theta_{\text{MP}} = arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\theta_{\text{MP}} = arg\max_{\theta} P(X|\theta)P(\theta)$$

Optimization problem!

MAP: problems

Not invariant to reparametrization



MAP: problems

Can't use as prior

$$P_k(\theta) = \frac{P(x_k|\theta)P_{k-1}(\theta)}{P(x_k)}$$

MAP: problems

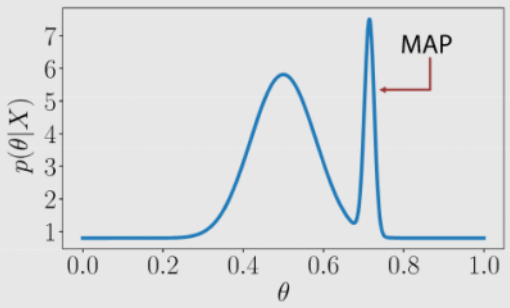
Can't use as prior

$$P_k(\theta) = \frac{P(x_k|\theta)P_{k-1}(\theta)}{P(x_k)}$$

$$P_k(\theta) = \frac{P(x_k|\theta)\delta(\theta - \theta_{\mathrm{MP}})}{P(x_k)} = \delta(\theta - \theta_{\mathrm{MP}})$$

MAP: problems

MAP is a solution to $L(\theta) = \mathbb{I}[\theta \neq \theta^*] \to \min_\theta$



MAP: problems

Objectives	Solution
$L(\theta) = \mathbb{I}[\theta \neq \theta^*] \to \min_\theta$	Mode
$L(\theta) = \mathbb{E}(\theta - \theta^*)^2 \to \min_\theta$	Mean
$L(\theta) = \mathbb{E}\lvert\theta - \theta^*\rvert \to \min_\theta$	Median

MAP: problems

Can't compute credible regions

$$\theta_{\text{MP}} = 12.53$$

MAP: problems

Can't compute credible regions

$$\theta_{\text{MP}} = 12.53$$

$$\theta_{\text{MP}} = 12.53 \pm 1000 \qquad \theta_{\text{MP}} = 12.53 \pm 0.001$$

$$\frac{8-1}{8+4-2} = \frac{7}{10}$$

Summary

Pros:

- Easy to compute

Cons:

- Not invariant to reparametrization
- Can't use as a prior
- Finds untypical point
- Can't compute credible intervals

conjugate

2. Conjugate distributions

Bayes formula

Fixed by model — Our own choice!

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Fixed by data

likelihood & posterior
lies in the same family of distributions

Conjugate prior

$P(\theta)$ is **conjugate** to $P(X|\theta)$:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$\mathcal{A}(v)$

$\mathcal{A}(v')$

Example

$$P(X|\theta) = \mathcal{N}(X|\theta, \sigma^2)$$

$$\mathcal{A}(v) = ?$$

$\mathcal{N}(X|\theta, \sigma^2)$ $\quad$ $\mathcal{A}(v)$

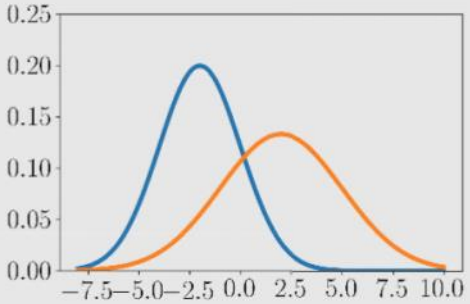
$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$\mathcal{A}(v')$

Two Gaussians

$$P(X_1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \qquad P(X_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

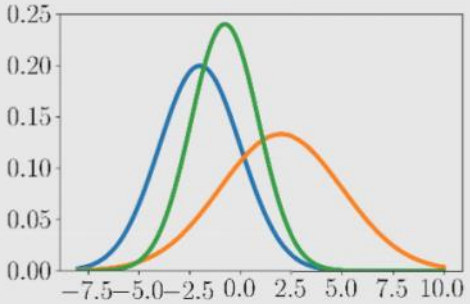
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \text{const} \cdot e^{-\text{parabola}}$$



Two Gaussians

$$P(X_1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \qquad P(X_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

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



Solution

$$P(X|\theta) = \mathcal{N}(X|\theta, \sigma^2)$$
$$\mathcal{A}(v) = \mathcal{N}(\theta|a, b^2)$$

$\mathcal{N}(X|\theta, \sigma^2)$ $\mathcal{N}(\theta|m, s^2)$

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$\mathcal{N}(\theta|a, b^2)$

Example

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\underline{\theta}, 1)\mathcal{N}(\theta|0, 1)}{p(x)}$$

Example

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\theta,1)\mathcal{N}(\theta|0,1)}{p(x)}$$

$$p(\theta|x) \propto e^{-\frac{1}{2}(x-\theta)^2} e^{-\frac{1}{2}\theta^2}$$

Example

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\theta,1)\mathcal{N}(\theta|0,1)}{p(x)}$$

$$p(\theta|x) \propto e^{-\frac{1}{2}(x-\theta)^2} e^{-\frac{1}{2}\theta^2}$$

$$p(\theta|x) \propto e^{-(\theta-\frac{x}{2})^2}$$

$p(\theta) \sim N(0, \sigma^2)$.

$p(x|\theta) \sim N(0, \sigma)$] Not conjugate!

Example

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\theta,1)\mathcal{N}(\theta|0,1)}{p(x)}$$

$$p(\theta|x) \propto e^{-\frac{1}{2}(x-\theta)^2} e^{-\frac{1}{2}\theta^2}$$

$$p(\theta|x) \propto e^{-(\theta-\frac{x}{2})^2}$$

$$p(\theta|x) = \mathcal{N}(\theta|\frac{x}{2}, \frac{1}{2})$$

conj-normal-gamma

Distributions: Gamma

Gamma distribution

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

Gamma distribution

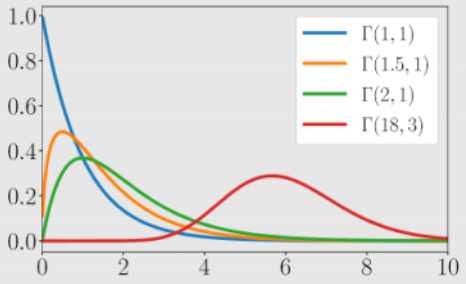
$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

$$\gamma, a, b > 0$$

Gamma distribution

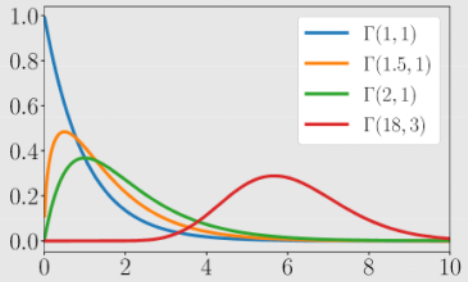
$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

$$\gamma, a, b > 0$$



Gamma distribution

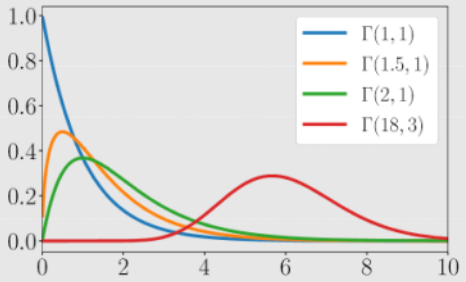
$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$



Gamma distribution

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

$$\Gamma(n) = (n-1)!$$

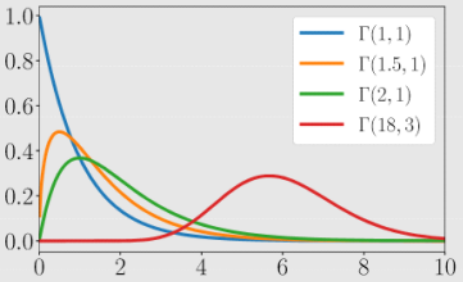


ТЕХНИЧЕСКИЙ СЛАЙД

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

$\Gamma(n) = (n-1)!$

Γ(5) = 24



Statistics

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

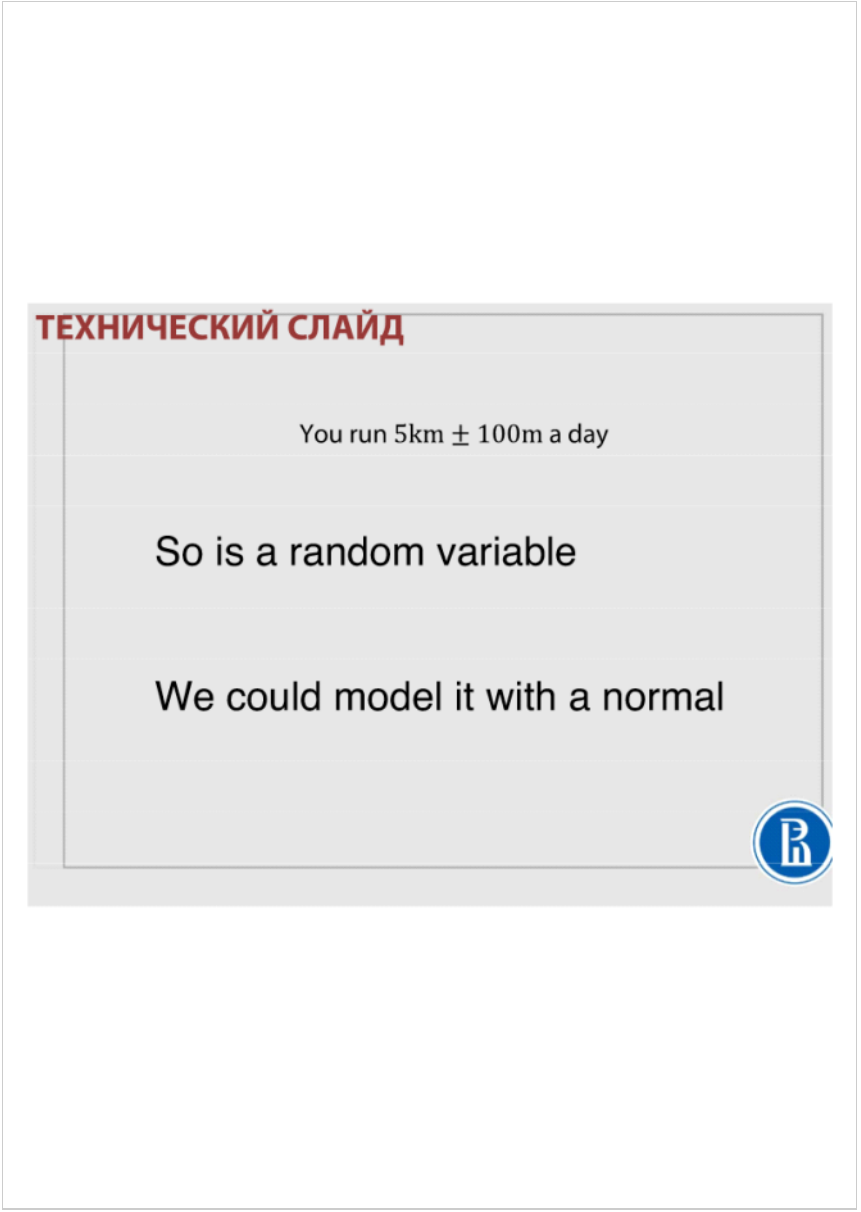
$$\mathbb{E}[\gamma] = a/b$$

$$\text{Mode}[\gamma] = \tfrac{a-1}{b}$$

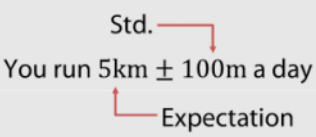
$$\text{Var}[\gamma] = a/b^2$$

Example

You run $5\text{km} \pm 100\text{m}$ a day


ТЕХНИЧЕСКИЙ СЛАЙД
You run 5km ± 100m a day
So is a random variable
We could model it with a normal

Example



Example

You run 5km $\pm$ 100m a day

$\mathbb{E}[x] = {}^{a}/_{b} = 5,\ Var[x] = {}^{a}/_{b^2} = 0.1^2$

Example

You run 5km $\pm$ 100m a day

$\mathbb{E}[x] = {}^{a}/_{b} = 5,\ Var[x] = {}^{a}/_{b^2} = 0.1^2$

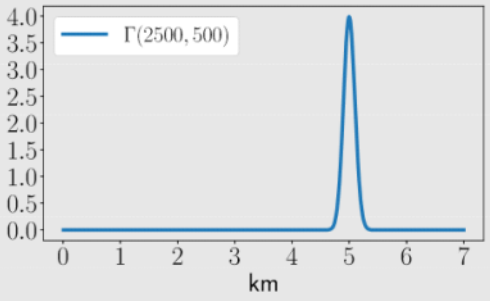
$\Rightarrow a = 2500, \mathrm{b} = 500$

Example

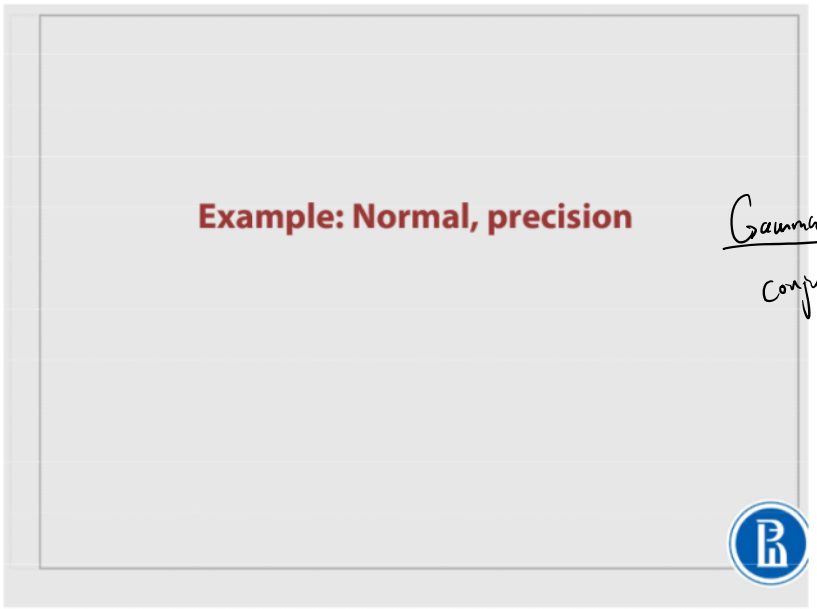
You run 5km $\pm$ 100m a day

$\mathbb{E}[x] = {}^{a}/_{b} = 5,\ Var[x] = {}^{a}/_{b^2} = 0.1^2$

$\Rightarrow a = 2500, \mathrm{b} = 500$



Example: Normal, precision

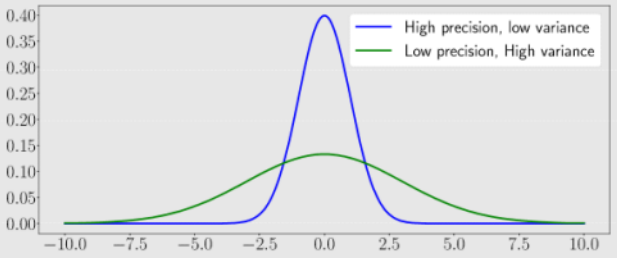


Gamma

Conjugate to normal dist wrt. precision

Precision

Precision $\gamma = \frac{1}{\sigma^2}$ Variance



Precision

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

Precision

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

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

Functional form

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

Functional form

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

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

Functional form

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

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

$$p(\gamma) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}?$$

Functional form

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

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

$$p(\gamma) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}?$$

$$p(\gamma|x) = \frac{p(x|\gamma)p(\gamma)}{p(x)} \propto \gamma e^{-\gamma(b+\frac{(x-\mu)^2}{2})}$$

Functional form

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

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

$$p(\gamma) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}?$$

$$p(\gamma|x) = \frac{p(x|\gamma)p(\gamma)}{p(x)} \propto \gamma e^{-\gamma(b+\frac{(x-\mu)^2}{2})}$$

Functional form

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

$$\mathcal{N}(x|\mu, \gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

$$p(\gamma) \propto \gamma^{\frac{1}{2}} e^{-b\gamma} ?$$

这个不 conjugate

$$p(\gamma|x) = \frac{p(x|\gamma)p(\gamma)}{p(x)} \propto \gamma e^{-\gamma(b + \frac{(x-\mu)^2}{2})}$$

Functional form

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

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

Functional form

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

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

$$p(\gamma) \propto \gamma^{a-1} e^{-b\gamma}$$

Functional form

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

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$$

$$p(\gamma) \propto \gamma^{a-1} e^{-b\gamma}$$

$$p(\gamma) = \Gamma(\gamma|a,b)$$

Gamma prior

$$p(\gamma) = \Gamma(\gamma|a, b) \propto \gamma^{a-1} e^{-b\gamma}$$

Gamma prior

$$p(\gamma) = \Gamma(\gamma|a,b) \propto \gamma^{a-1}e^{-b\gamma}$$

$$p(\gamma|x) \propto p(x|\gamma)p(\gamma)$$

Gamma prior

$$p(\gamma) = \Gamma(\gamma|a,b) \propto \gamma^{a-1}e^{-b\gamma}$$

$$p(\gamma|x) \propto p(x|\gamma)p(\gamma)$$

$$p(\gamma|x) \propto \left(\gamma^{\frac{1}{2}}e^{-\gamma\frac{(x-\mu)^2}{2}}\right) \cdot \left(\gamma^{a-1}e^{-b\gamma}\right)$$

Gamma prior

$$p(\gamma) = \Gamma(\gamma|a,b) \propto \gamma^{a-1}e^{-b\gamma}$$

$$p(\gamma|x) \propto p(x|\gamma)p(\gamma)$$

$$p(\gamma|x) \propto \left(\gamma^{\frac{1}{2}}e^{-\gamma\frac{(x-\mu)^2}{2}}\right) \cdot \left(\gamma^{a-1}e^{-b\gamma}\right)$$

$$p(\gamma|x) \propto \gamma^{\frac{1}{2}+a-1}e^{-\gamma(b+\frac{(x-\mu)^2}{2})}$$

Gamma prior

$$p(\gamma) = \Gamma(\gamma|a,b) \propto \gamma^{a-1}e^{-b\gamma}$$

$$p(\gamma|x) \propto p(x|\gamma)p(\gamma)$$

$$p(\gamma|x) \propto \left(\gamma^{\frac{1}{2}}e^{-\gamma\frac{(x-\mu)^2}{2}}\right) \cdot \left(\gamma^{a-1}e^{-b\gamma}\right)$$

$$p(\gamma|x) \propto \gamma^{\frac{1}{2}+a-1}e^{-\gamma(b+\frac{(x-\mu)^2}{2})}$$

$$p(\gamma|x) = \Gamma(a+\frac{1}{2}, b+\frac{(x-\mu)^2}{2})$$

conjugate-beta

Distributions: Beta

Beta distribution

$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

Beta distribution

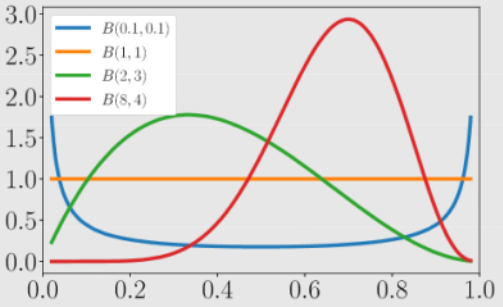
$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$x \in [0,1], \quad a,b > 0$$

Beta distribution

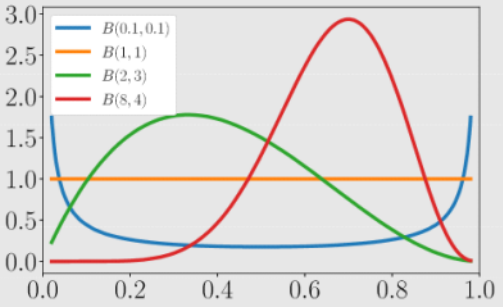
$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$x \in [0,1], \quad a,b > 0$$



Beta distribution

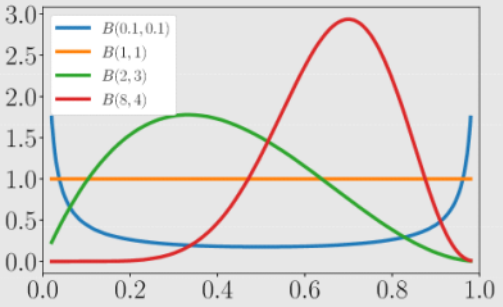
$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$



Beta distribution

$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$



Statistics

$$B(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$\mathbb{E}x = \frac{a}{a+b}$$
$$\text{Mode}[x] = \frac{a-1}{a+b-2}$$
$$\text{Var}[x] = \frac{ab}{(a+b)^2(a+b-1)}$$

Example

Movie rank is 0.8 ± 0.1

Example ТЕХНИЧЕСКИЙ СЛАЙД

Movie rank is 0.8 ± 0.1

1 — best movie
0 — Batman & Robin

Example

Movie rank is 0.8 ± 0.1

$$\mathbb{E}x = \frac{a}{a+b} = 0.8$$

$$\mathrm{Var}[x] = \frac{ab}{(a+b)^2(a+b-1)} = 0.1^2$$

Example

Movie rank is 0.8 ± 0.1

$$\mathbb{E}x = \frac{a}{a+b} = 0.8$$

$$\text{Var}[x] = \frac{ab}{(a+b)^2(a+b-1)} = 0.1^2$$

$$\Rightarrow a = 12, b = 3$$

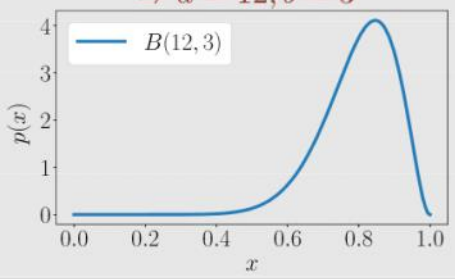
Example

Movie rank is 0.8 ± 0.1

$$\mathbb{E}x = \frac{a}{a+b} = 0.8$$

$$\text{Var}[x] = \frac{ab}{(a+b)^2(a+b-1)} = 0.1^2$$

$$\Rightarrow a = 12, b = 3$$



Example: Bernoulli

Beta prior

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Beta prior

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Beta prior

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

Beta prior

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

$$p(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0} \cdot \theta^{a-1}(1-\theta)^{b-1}$$

Beta prior

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

$$p(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0} \cdot \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

Beta prior

$$p(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

$$p(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0} \cdot \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

$$p(\theta|X) = B(N_1+a, N_0+b)$$

Summary

Summary

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Summary

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Summary

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Summary

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Pros and cons

Pros:

Pros and cons

Pros:

- Exact posterior

Pros and cons

Pros:

- Exact posterior
- Easy for on-line learning

E.g. $p(\theta|X) = B(N_1 + a, N_0 + b)$

Pros and cons

Pros:

- Exact posterior
- Easy for on-line learning

E.g. $p(\theta|X) = B(N_1 + a, N_0 + b)$

Cons:

Pros and cons

Pros:

- Exact posterior
- Easy for on-line learning

E.g. $p(\theta|X) = B(N_1 + a, N_0 + b)$

Cons:

- Conjugate prior may be inadequate

→ Next week.
what if no conjugate posterior
full posterior / approx. posterior

3. Choose correct statements:

- [ ] $p(a \mid b, c) = p(a \mid b)p(a \mid c)$ when b and c are independent

  Un-selected is correct

- [x] $p(a \mid b, c) = \frac{p(b \mid a, c)p(a \mid c)}{\int p(b \mid a', c)p(a' \mid c)da'}$

  Correct

- [ ] $p(a \mid b) = \frac{p(a, c \mid b)}{p(c \mid a, b)}$

  This should be selected

- [x] $p(a \mid b)p(b) + p(a \mid \overline{b})p(\overline{b}) = p(a)$, for binary b

  Correct
  The law of total probability

- [ ] $p(a \mid b) + p(a \mid \overline{b}) = p(a)$, for binary b

  Un-selected is correct

2. Choose correct statements:

- [x] $p(a \mid b) = \int p(a \mid b, c)p(c)dc$, when b and c are independent

  Correct
  $p(c \mid b) = p(c)$ when c and b are independent

- [ ] $p(a \mid b) = \int p(a \mid b, c)dc$

  Un-selected is correct

- [ ] $p(a \mid b) = \int p(a, c \mid b)dc$

  This should be selected

- [x] $p(a, b \mid c) = p(a \mid b, c)p(b \mid c)$

  Correct

4. Let joint probability over random variables a, b, c be $p(a, b, c) = p(a|b)p(b|c)p(c)$. Are random variables a and c independent?

- [ ] Yes
- [x] No

Correct
Let's marginalize joint probability by b and we get $\int p(a, b, c)db = \int p(a|b)p(b|c)p(c)db = p(c) \int p(a|b)p(b|c)db$. Unfortunately, integral contain inside both a and c and it can't be decomposed into two integrals $\int f(a, b)db$ and $\int g(c, b)db$, so a and c is dependent.

5. Let joint probability over random variables a, b, c, d be $p(a, b, c, d) = p(a|b)p(b)p(c|d)p(d)$. Are random variables a and c independent?

- [x] Yes

Correct
Let's marginalize joint probability by b and d, so we get $\int p(a, b, c, d)\, db\, dd = \int p(a|b)p(b)p(c|d)p(d)\, db\, dd =$ $= \left(\int p(a|b)p(b)\, db\right) \left(\int p(c|d)p(d)\, dd\right)$. So we decomposed it into two integrals $\int f(a, b)db$ and $\int g(c, d)dd$, so a and c is independent.

- [ ] No