# NPFL108 – Bayesian inference

# Introduction

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek

Version: 21/02/2014

# Outline

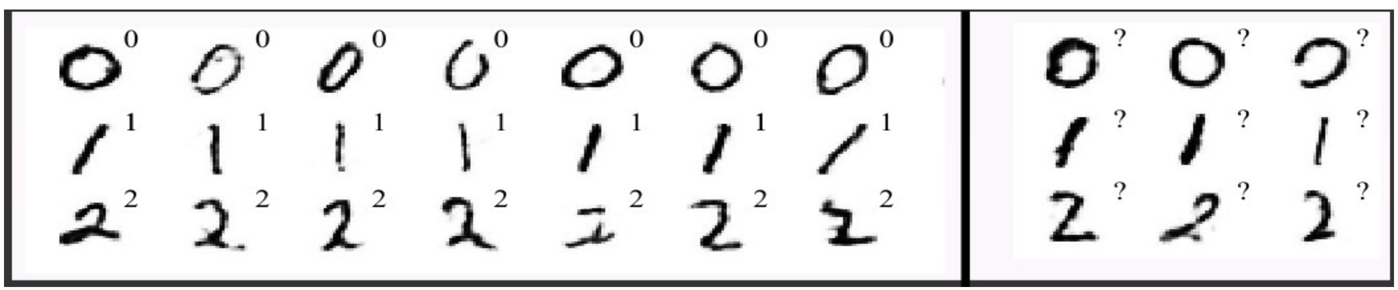
- The course objective
- Syllabus
- Literature
- Course structure
- Basics of Probability Theory
- Basic probability distributions

# The course objective

- The course aims to provide students with basic understanding of modern Bayesian inference methods used in Bayesian Machine Learning.
- Being Bayesian is about managing uncertainty and efficient use of data.
- In many tasks such as stock trading or speech recognition, the uncertainty is inherent and there is always less data then we really need.

# What is Machine Learning?

- The design of computational systems that discover patterns in a collection of data instances in an automated manner.
- The ultimate goal is to use the discovered patterns to make predictions on new data instances not seen before.

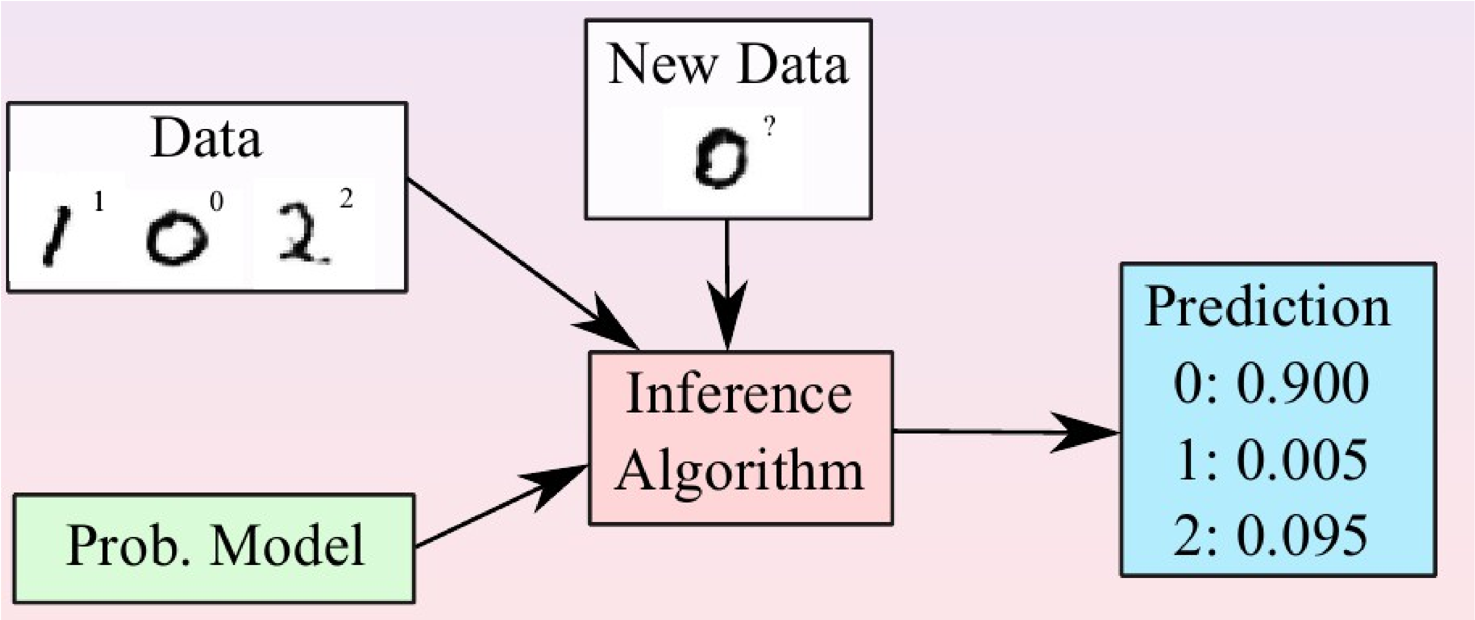


- Instead of manually encoding patterns in computer programs, we make computers learn these patterns without explicitly programming them.

Figure source [Hinton et al. 2006].

# Model-based Machine Learning

- We design a probabilistic model which explains how the data is generated.
- An inference algorithm combines model and data to make predictions.
- Probability theory is used to deal with uncertainty in the model or the data.



# Syllabus #1

- Introduction
  - Random variables
  - Sum rule, product rule, Bayes rule
  - Independence
  - Prior, likelihood, posterior, predictions
  - Basic probability distributions
- Types of priors
  - Conjugate prior vs. Non-conjugate prior
  - Proper prior vs. improper prior
  - Informative prior vs. uninformative prior

# Syllabus #2

- Bayesian inference for parameters of the normal distribution
  - Unknown mean, known variance
  - Known mean, unknown variance
  - Unknown mean and variance, conjugate prior
  - Unknown mean and variance, non-conjugate prior

# Syllabus #3

- Inference in discrete graphical models
  - Variables, Parameters, Networks, Plate notation
  - Conditional Independence
  - Markov blanket
  - Message passing, Belief propagation, Loopy belief propagation
- Approximate Inference
  - Variational Inference / Bayes
  - Expectation propagation
  - Sampling methods
    - Metropolis-Hastings, Gibbs, slice sampling, random walk
- Non-parametric Bayesian Methods
  - Gaussian processes
  - ~~Dirichlet processes~~

# Paradigm shift

- Point estimates vs. posterior estimates
- An example:
  - flip of coins (data): H T H H
  - point estimate 3/4
  - Bayesian estimate ?

  - flip of coins (data): H T H H T H H H
  - point estimate 3/4
  - Bayesian estimate ?
  - We need distributions over parameters!

# True or False

- \- being Bayesian is just about having priors
- \+ being Bayesian is about managing uncertainty

- \- Bayesian methods are slow
- \+ Bayesian methods can be as fast as Expectation-Maximisation

- \- Non parametric means no parameters
- \+ Non parametric means the number of parameters grows as necessary given the data

- \- Variational inference is complicated
- \+ Variational inference is an extension of Expectation-Maximization

# Literature


- C. M. Bishop, Pattern Recognition and Machine Learning, vol. 4, no. 4. Springer, 2006, p. 738.
- MacKay, David JC. Information theory, inference and learning algorithms. Cambridge university press, 2003.
- Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.
- B. Thomson and S. Young, Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems," Computer Speech and Language, vol. 24, no. 4, pp. 562-588, 2010.
- D. Marek (studijní obor teoretická informatika): Implementace aproximativních Bayesovských metod pro odhad stavu v dialogových systémech, Dimplomová práce, UFAL, MFF, CUNI, 2013.

# Course structure

- Mixed lectures and practicals
- Each of you will have its own lecture:
  - description of some inference problem
  - derivation of the solution
  - presentation of implementation of the problem in some programming language
    - I prefer Python ;-) using only NumPy and SciPy
  - A lot of work!

# Available problems

- **You can invent your own**
- Message passing, belief propagation, loopy belief propagation in discrete graphical models
- Laplace approximation: The probit regression model
- Variational inference: The probit regression model
- Variational inference: 2D Ising Model
- Variational inference: Unknown mean and variance of a Gaussian with improper priors.
- Expectation Propagation: The Clutter Problem
- Variational inference: The Clutter Problem
- Expectation Propagation: The probit regression model
- Bayesian inference for regression
- Gibbs Sampling: Probit regression
- Metropolis-Hastings Random walk: Logit regression
- Gibbs Sampling: Probit regression in multi-class setting
- Gaussian Processes: Sampling from a GP, inference GP with prior m(x) = 0 and k(x,x') = {squared exponential, ration quadratic}, analyse impact of different covariance functions of the resulting approximations.

# Basics of Probability Theory

- Random variables
- Sum rule, product rule, Bayes rule
- Independence
- Prior, likelihood, posterior, predictions

# Random variable

- Random variable (RV) is a variable whose value is subject to variations due to chance (i.e. randomness, in a mathematical sense)
- A random variable conceptually does not have a single, fixed value
- It can take on a set of possible different values, each with an associated probability
- We talk about a probability distribution for values of some RV

$$P(X)$$

# Examples of random variables

- rolling a die (head or tail)
- person's marriage status (no, yes)
- person's number of children (0, 1, 2, ...)
- person's height (real numbers between 0 and +inf)
- temperature the next year the same day

- parameters of the distributions of describing another RVs

- Basic division of RV is:
  - Discrete
  - Continuous

# Sum rule #1

- Let's have two RVs:
  - X - person's marriage status
  - Y - person's number of children
- Then
  - P(X=yes) is the probability that a person is married
  - P(Y=0) is the probability that a person has exactly one child
  - P(X=yes, Y=0) is the **JOINT** probability that a person is married and has exactly one child
  - P(X=yes | Y=0) is the **CONDITIONAL** probability that a person is married if we know that he/she has exactly one child

# Sum rule #2

- Computes marginal probabilities from joint probabilities.
- Sum rule says:

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

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

- In more precise notation:

$$P(X = x_i) = \sum_{y_j} P(X = x_i, Y = y_j)$$

# Product rule

- Relates joint probability with conditional and marginal probability marginal probability.
- Product rule says:

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

# Bayes rule

- The theory of probability can be derived using just sum and product rules
- Bayes' theorem gives the relationship between the probabilities of X and Y and the conditional probabilities of X given Y and Y given Z

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

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

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

# Independence

- Independence of X and Y:

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

- Conditional independence of X and Y given Z:

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

# Bayesian Model Framework

- The probabilistic model M with parameters θ explains how the data D is generated by specifying the likelihood function p(D|θ, M).
- Our initial uncertainty on θ is encoded in the prior distribution p(θ|M).
- Bayes' rule allows us to update our uncertainty on θ given D (posterior):

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

# Bayesian predictions

- We can then generate probabilistic predictions for some new data point x given D and M using:

$$P(x|D,M)=\int P(x|\theta)\,P(\theta|D,M)\,d\theta$$

- Example: Buttered toast phenomenon

# Probabilistic Graphical Models

- The Bayesian framework requires to specify a high-dimensional distribution $p(x_1, \ldots, x_k)$ on the data, model parameters and latent variables.
- Working with fully flexible joint distributions is intractable!
- We will work with structured distributions, in which the random variables interact directly with only few others. These distributions will have many conditional independences.
- This structure will allow us to:
  - Obtain a compact representation of the distribution.
  - Use computationally efficient inference algorithms.
- The framework of probabilistic graphical models allows us to represent and work with such structured distributions in an efficient manner.

# Examples of Probabilistic Graphical Models

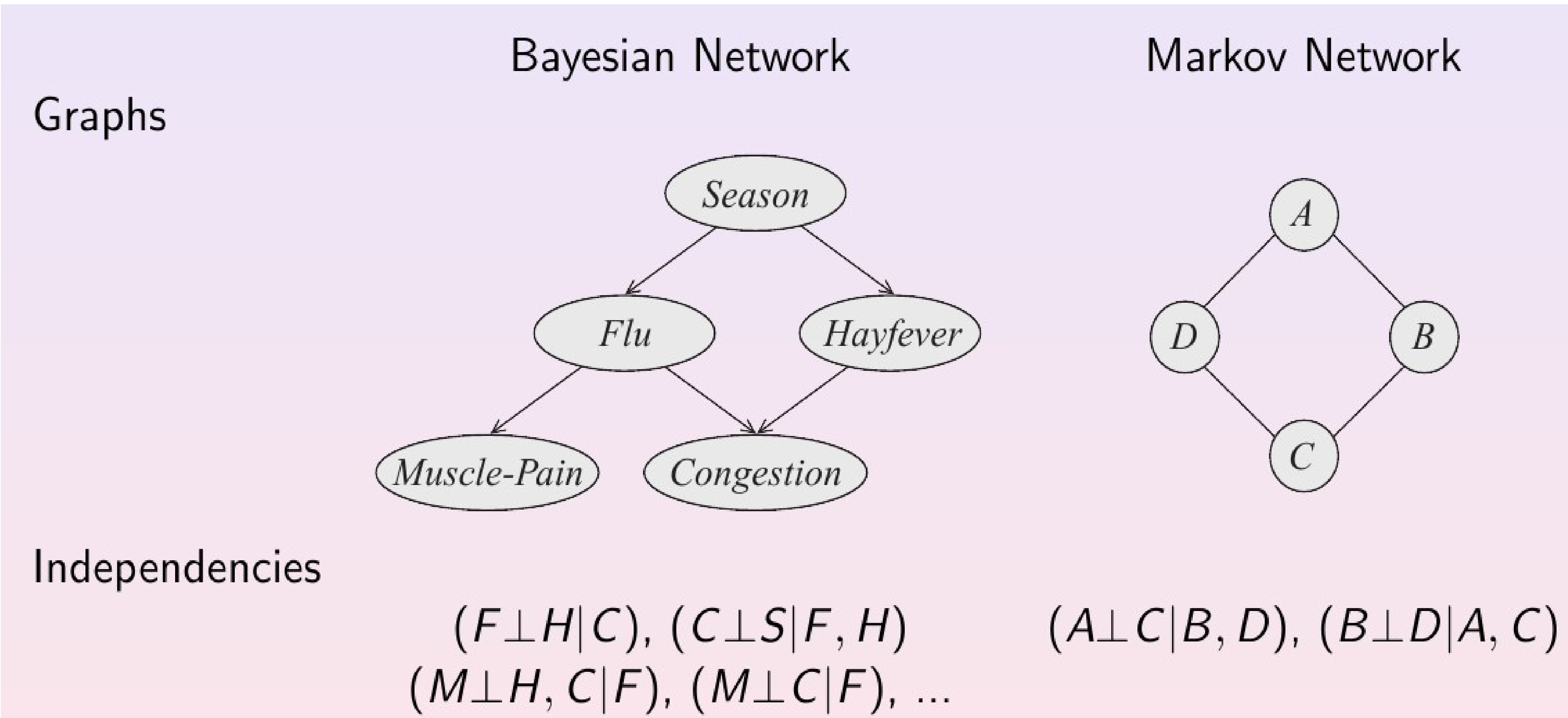



Figure source [Koller et al. 2009].

# Examples of Probabilistic Graphical Models

BN Examples: Naive Bayes

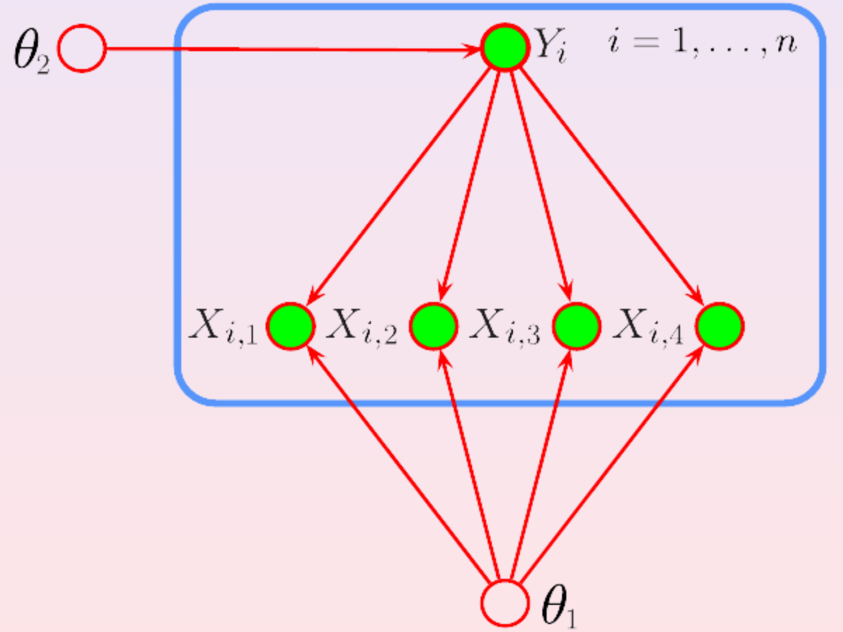



Figure source [Murphy 2012].

BN Examples: Hidden Markov Model

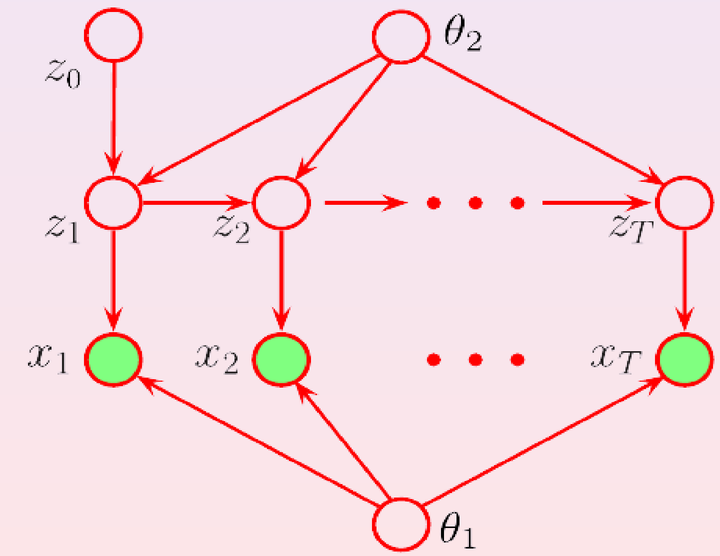



Figure source [Murphy 2012].

# Basic probability distributions

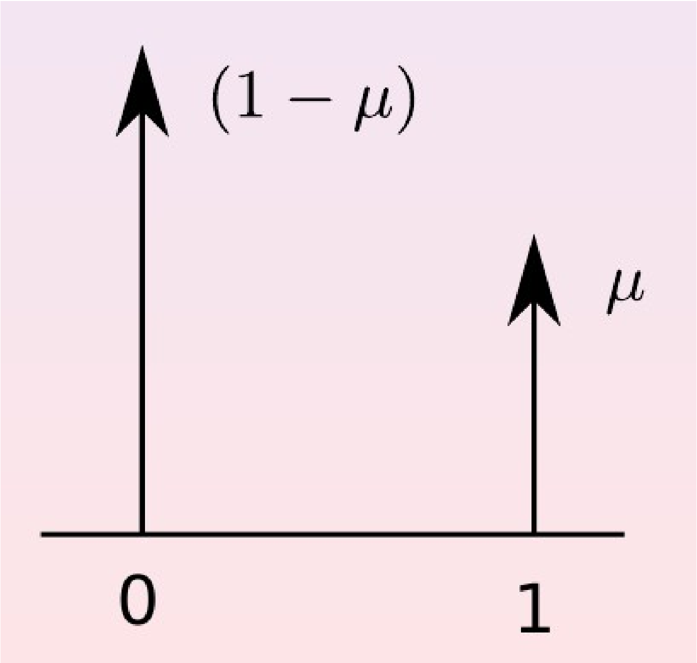
- Bernoulli, Binomial
- Beta
- Categorical (Discrete), Multinomial
- Dirichlet
- (Multivariate) Normal
- Gamma and inverse gamma
- Wishart

# Bernoulli

- Distribution for $x \in \{0, 1\}$ governed by $\mu \in [0, 1]$ such that $\mu = p(x = 1)$.

$$Bern(x; \mu) = \mu^x (1 - \mu)^{1-x}$$

  - $E(x) = \mu$
  - $Var(x) = \mu(1 - \mu)$

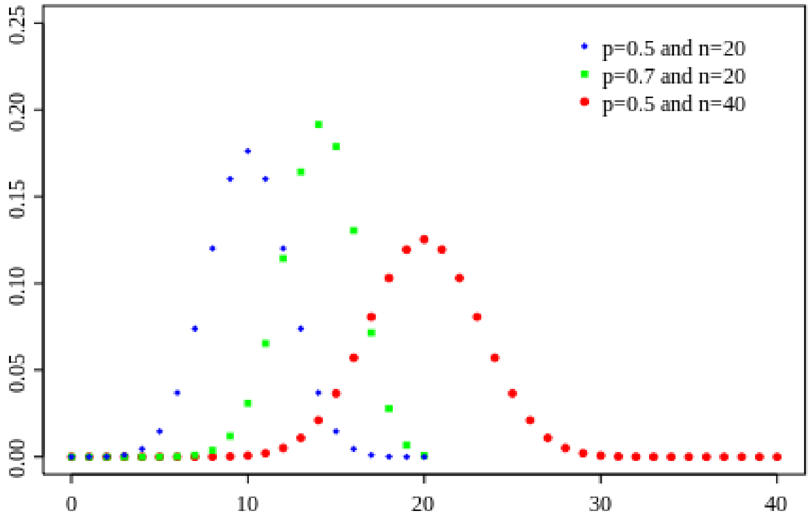


# Binomial

- Binomial distribution is a variation of Bernoulli distribution for multiple trials.
- Distribution for m ∈ {0, 1, .., N} governed by μ ∈ [0, 1] probability of success in N trials.

$$Bin(m; N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- E(m) = Nμ
- Var(m) = Nμ(1 − μ).




Source Wikipedia

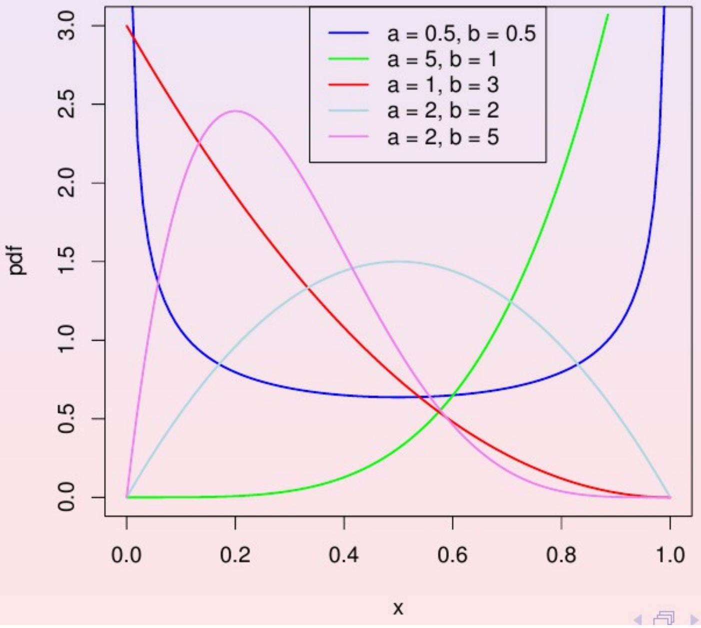
# Beta

- Distribution for μ ∈ [0, 1] such as the probability of a binary event.

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}.$$

$$\mathbb{E}(x) = a/(a+b).$$

$$\mathrm{Var}(x) = ab/((a+b)^2(a+b+1)).$$




- Beta is very often used as a prior for parameters of Bernoulli and Binomial distributions

# Multinomial

- We extract with replacement n balls of k different categories from a bag.
- Let $x_i$ and denote the number of balls extracted and $p_i$ the probability, both of category i = 1, . . . , k
  - e.g. {0,0,5,2,4,0,0}

$$p(x_1, \ldots, x_k | n, p_1, \ldots, p_k) = \begin{cases} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} & \text{if} \quad \sum_{i=1}^{k} x_k = n \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}(x_i) = np_i.$$

$$\mathrm{Var}(x_i) = np_i(1 - p_i).$$

$$\mathrm{Cov}(x_i, x_j) = -np_i p_j(1 - p_i).$$

# Dirichlet

Multivariate distribution over $\mu_1, \ldots, \mu_k \in [0, 1]$, where $\sum_{i=1}^{k} \mu_i = 1$.

Parameterized in terms of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ with $\alpha_i > 0$ for $i = 1, \ldots, k$.

$$\mathrm{Dir}(\mu_1, \ldots, \mu_k | \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_k\right)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \prod_{i=1}^{k} \mu_i^{\alpha_i} .$$

$$\mathbb{E}(\mu_i) = \frac{a_i}{\sum_{j=1}^{k} a_j} .$$

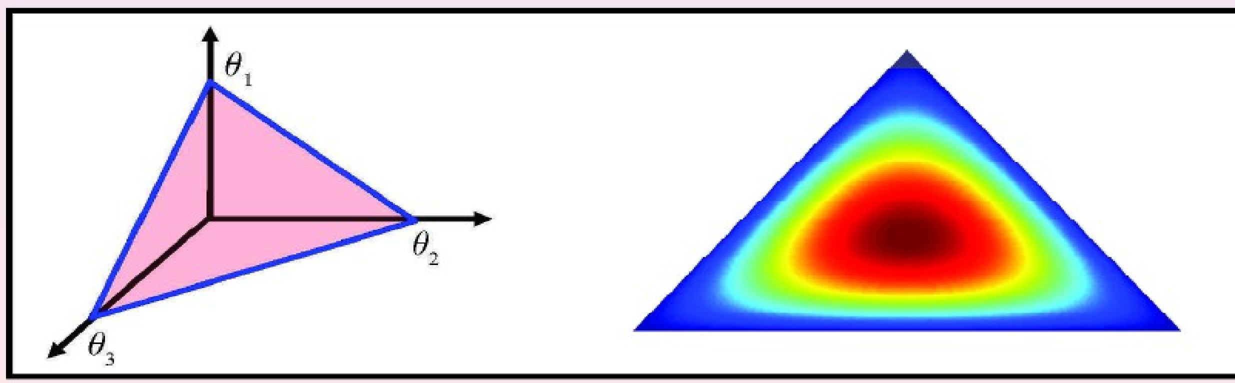



Figure source [Murphy 2012].

- Dirichlet is very often used as a prior for parameters of a Multinomial distribution

# Multivariate Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$.

$\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$.

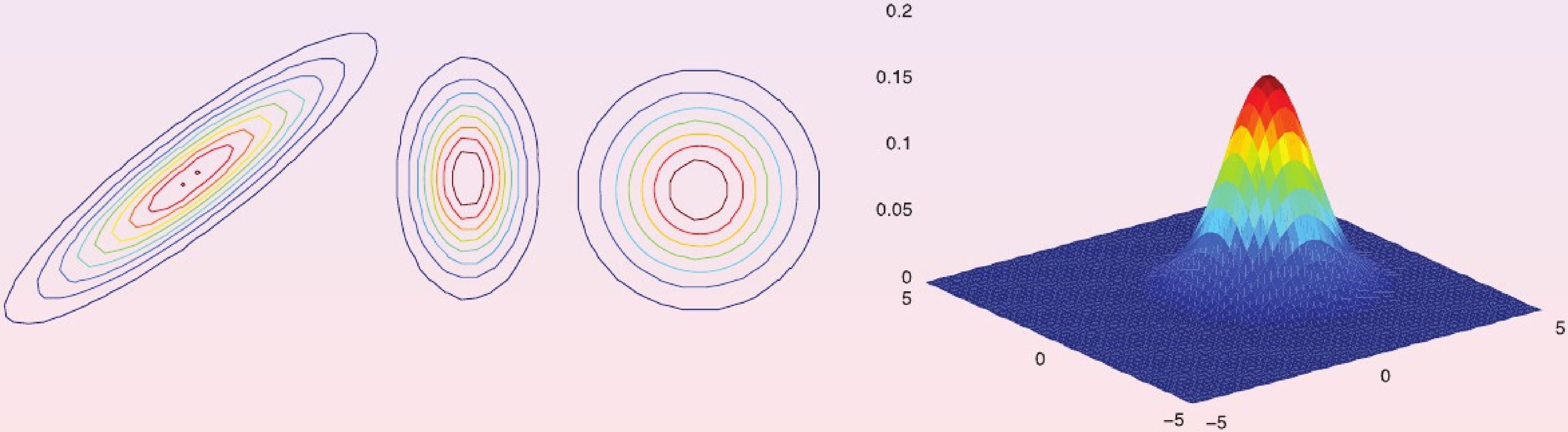



Figure source [Murphy 2012].

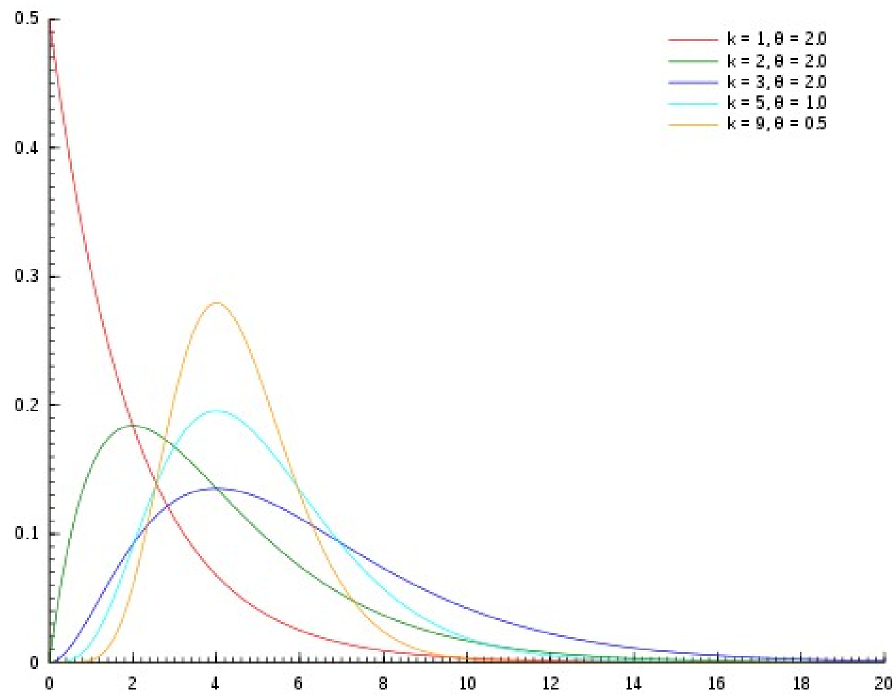
# Gamma

- Distribution for $\tau > 0$ governed by a > 0 and b > 0

$$Gam(\tau; a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}$$

  - $E(\tau) = a/b$
  - $Var(\tau) = a/b^2$

- Gamma is very often used as a prior for a precision of a normal distribution
- Inverse Gamma is typically used as a prior for variance




Source Wikipedia

# Wishart

- Wishart is an multivariate equivalent of gamma distribution.

Distribution for the precision matrix $\Lambda = \mathbf{\Sigma}^{-1}$ of a Multivariate Gaussian.

$$\mathcal{W}(\Lambda|\mathbf{w}, \nu) = B(\mathbf{W}, \nu)|\Lambda|^{(\nu-D-1)} \exp\left\{-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\Lambda)\right\} ,$$

where

$$B(\mathbf{W}, \nu) \equiv |\mathbf{W}|^{-\nu/2} \left(2^{\nu D/2}\pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu+1-i}{2}\right)\right) .$$

$\mathbb{E}(\Lambda) = \nu\mathbf{W}$.

# Summary

- With ML computers learn patterns and then use them to make predictions.
- With ML we avoid to manually encode patterns in computer programs.
- Model-based ML separates knowledge about the data generation process (model) from reasoning and prediction (inference algorithm).
- The Bayesian framework allows us to do model-based ML using probability distributions which must be structured for tractability.
- Probabilistic graphical models encode such structured distributions by specifying several CIs (factorizations) that they must satisfy.

# Thank you!

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek