# Nonparametric Bayesian Methods and Clustering


Mikhail Karasikov

https://bmi.inf.ethz.ch/


11 – 13 December 2019

# Tutorial Outline

1. Nonparametric models
2. Bayesian nonparametrics
3. Clustering

# Literature


- Peter Orbanz, *Lecture Notes on Bayesian Nonparametrics*. (2014)
- Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag (2006)
- Trevor Hastie, Robert Tibshirani & Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag (2001)
- Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press (2012) $\rightarrow$ Chapter 6
- Richard O. Duda, Peter E. Hart & David G. Stork, *Pattern Classification*. Wiley & Sons (2001)
- Vladimir N. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer Verlag (1983)
- Ulf Grenander, General Pattern Theory: A Mathematical Study of Regular Structures. Oxford University Press (1993)
- Andrew Webb, *Statistical Pattern Recognition*. Wiley & Sons, (2002)
- Keinosuke Fukunaga, *Statistical Pattern Recognition*. Academic Press (1990)
- Brian D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press (1996)
- Larry Wasserman, *All of Statistics*. (1st ed. 2004. Corr. 2nd printing) Springer Verlag (2004)

1. Nonparametric models

2. Bayesian nonparametrics

3. Clustering

# Parametric vs Nonparametric

Consider a *statistical model*

$$\mathcal{M} = \{P_\theta(x) \mid \theta \in \Theta\}$$

— a set of probability measures on $\mathbb{X}$ indexed by a parameter $\theta$.

## Definition

A model $\mathcal{M}$ is called *nonparametric* if its parameter space has infinite dimension.

## Examples

- Usually in ML we have a finite number of parameters $\Theta \equiv \mathbb{R}^n$ and work with *parametric* models.
  (e.g., finite mixture of Gaussian distributions)
- Now consider the space of all continuous functions on $\mathbb{R}^m$.
  The dimension is infinite, hence the model is *nonparametric*.
  Another example: infinite mixtures of Gaussians

1. Nonparametric models

2. Bayesian nonparametrics

3. Clustering

# Bayesian approach

Deriving the posterior in Bayesian inference

1. Set a prior $\pi(\theta)$ on parameters
2. Define a parametric model $p(x|\theta)$
3. Compute the posterior $\pi(\theta|\mathbf{X})$ using the Bayes rule

$$\pi(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)\pi(\theta)}{p(\mathbf{X})}$$

Now, consider a **nonparametric** model $\mathcal{M} = \{P_\theta(x) \mid \theta \in \Theta\}$.

The parameter space $\Theta$ has infinite dimension
$\implies$ Lebesgue measure and integral cannot be defined[1]
$\implies$ Issues with working with prior and posterior
$\implies$ **Need a different approach**

[1] the probability measure, however, may be still well defined

# Nonparametric Bayesian methods

We want to define distributions on an infinite-dimensional space $\Theta$

- How to construct a prior?
- How to compute the posterior?

**Approach:** Use *stochastic processes* and draw samples from them

- Define an algorithm for drawing from the prior
- Construct an algorithm for drawing from the posterior

**A common prior distribution:** *Dirichlet Process*

# The Dirichlet process

## Parameters

- Base distribution $F_0$
- Concentration parameter $\alpha \in (0, \infty)$

## Sampling algorithm

1. Draw the first sample: $X_1 \sim F_0$.
2. For $i = 2, 3, \ldots$, draw

$$X_i | X_{i-1}, \ldots, X_1 = \begin{cases} X \sim \hat{F}_{i-1}, & \text{with probability } p = \frac{i-1}{\alpha+i-1} \\ X \sim F_0, & \text{with probability } p = \frac{\alpha}{\alpha+i-1} \end{cases}$$

where $\hat{F}_{i-1}$ is the empirical distribution of $X_1, \ldots, X_{i-1}$.

# Exercise

Find the asymptotics of the number of distinct samples drawn in the DP sampling algorithm

$$E\left[\sum_{i=1}^{n} 1\{X_i \text{ is drawn from } F_0\}\right] \sim f(n).$$

**Find** $f(n)$

1. Nonparametric models

2. Bayesian nonparametrics

3. Clustering

# Mixture models

**Model:** mixture distribution

$$p(x) = \sum_{k=1}^{K} c_k p(x|\theta_k)$$

**Maximum likelihood estimation:** with the EM algorithm

# Dirichlet process mixture model

**Prior** is given as a DP

$$\theta \sim \mathsf{DP}(\alpha, F_0), \quad F_0 \text{ is usually Gaussian: } \mathcal{N}(\mu, \Sigma)$$

**Model:** infinite mixture distribution

$$p(x) = \sum_{k=1}^{\infty} c_k p(x|\theta_k)$$

**Drawing from the posterior:** with the Gibbs sampling algorithm