Statistical Machine Learning

Week 01 – Lrcture 01 (Version 1.0)

Non-Parametric Bayesian Models

Hamid R. Rabiee

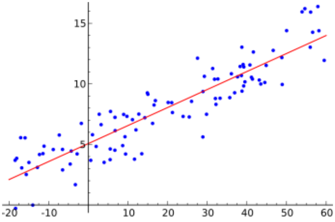
Spring 2023

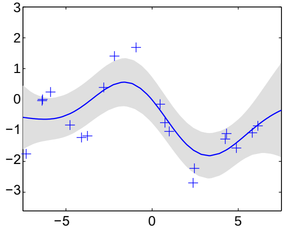
Acknowledgement: Contents from P. Orbanz & Z. Ghahremani.

Parameters and Patterns

Parameters

$$P(X/\theta) \quad = \quad \text{Probability[data/pattern]}$$





Inference idea

$$\text{data} \quad = \quad \text{underlying pattern} \quad + \quad \text{independent noise}$$

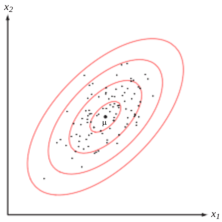
Parametric model

- Number of parameters fixed (or constantly bounded) w.r.t. sample size

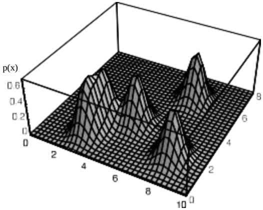
Nonparametric model

- Number of parameters grows with sample size
- ∞-dimensional parameter space

Example: Density estimation



Parametric



Nonparametric

NONPARAMETRIC BAYESIAN MODEL

Definition

A nonparametric Bayesian model is a Bayesian model on an ∞-dimensional parameter space.

Interpretation

Parameter space $\mathcal{T}$ = set of possible patterns, for example:

Problem	$\mathcal{T}$
Density estimation	Probability distributions
Regression	Smooth functions
Clustering	Partitions

Solution to Bayesian problem = posterior distribution on patterns

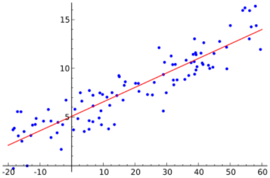
Can we justify our assumptions?

Recall:

data = pattern + noise

In Bayes' theorem:

$$Q(d\theta|x_1,\ldots,x_n) = \frac{\prod_{j=1}^{n} p(x_j|\theta)}{p(x_1,\ldots,x_n)} Q(d\theta)$$



Definition

$X_1, X_2, \ldots$ are *exchangeable* if $P(X_1, X_2, \ldots)$ is invariant under any permutation σ:

$$P(X_1 = x_1, X_2 = x_2, \ldots) = P(X_1 = x_{\sigma(1)}, X_2 = x_{\sigma(2)}, \ldots)$$

In words:

Order of observations does not matter.

De Finetti's Theorem

$$P(X_1 = x_1, X_2 = x_2, \ldots) = \int_{\mathbf{M}(\mathcal{X})} \Big(\prod_{j=1}^{\infty} \theta(X_j = x_j)\Big) Q(d\theta)$$

$$\Updownarrow$$

$$X_1, X_2, \ldots \text{ exchangeable}$$

where:

- $\mathbf{M}(X)$ is the set of probability measures on X
- θ are values of a random probability measure Θ with distribution Q

Implications

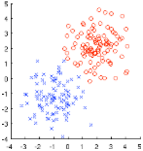
- Exchangeable data decomposes into pattern and noise
- More general than i.i.d.-assumption
- Caution: θ is in general an ∞-dimensional quantity

Why Nonparametric?

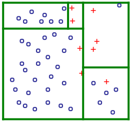
Please Pay Attention!

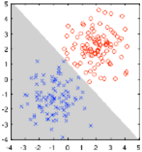
Nonparametric Does NOT mean there are no parameters.

Example: Classification

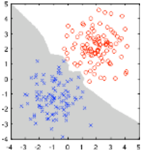


Data

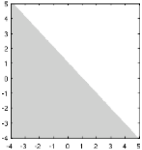




Build model



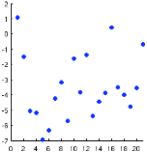
Nonparametric Approach



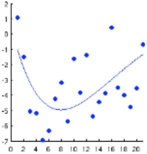
Predict using model

Parametric Approach

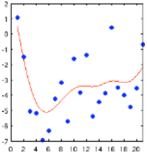
Example: Regression



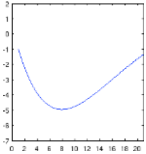
Data



Build model



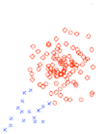
Nonparametric Approach



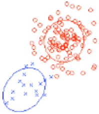
Predict using model

Parametric Approach

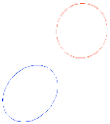
Example: Clustering

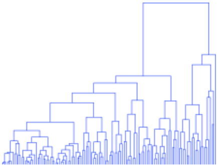


Data



Build model





Nonparametric Approach

Parametric Approach

Why Bayesian?

Why Bayesian?

You can take a course on this question. One answer:

Infinite Exchangeability: $\forall n \; p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$

De Finetti's Theorem (1955): If $(x_1, x_2, \ldots)$ are *infinitely exchangeable*, then $\forall n$

$$p(x_1, \ldots, x_n) = \int \left(\prod_{i=1}^{n} p(x_i|\theta) \right) dP(\theta)$$

for some random variable θ.

Why Bayesian: Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is?

The maximum likelihood estimate is $\theta = ?$. Seems reasonable?

Why Bayesian: Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Now suppose we observe: $\{$H, H, H, H$\}$. What do we think θ is?

The maximum likelihood estimate is $\theta = ?$. Seems reasonable?

Why Bayesian: Simple Example

When we observe $\{H, H, H, H\}$, why does $\theta = 1$ seem unreasonable?

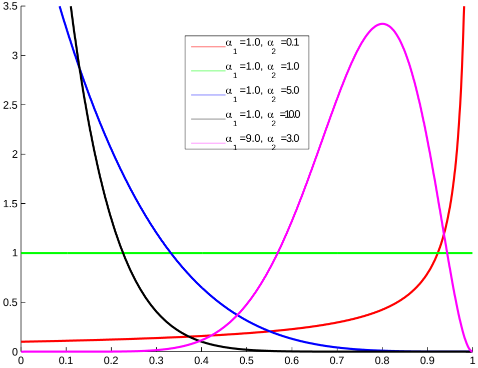
Prior knowledge! We believe coins generally have $\theta \approx 1/2$. How to encode this? By using a Beta *prior on* θ.

Bayesian Approach to Estimating θ

Place a Beta(a, b) prior on θ. This prior has the form:

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

What does this distribution look like?

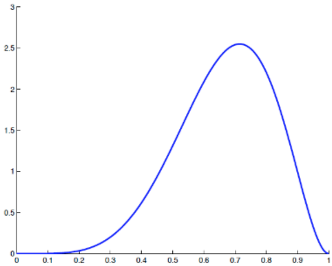


Bayesian Approach to Estimating θ

After observing X, a sequence with n heads and m tails, the posterior on θ is:

$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)p(\theta) \\ &\propto \theta^{a+n-1}(1-\theta)^{b+m-1} \\ &\sim \mathrm{Beta}(a+n, b+m). \end{aligned}$$

If $a = b = 1$ and we observe 5 heads and 2 tails, Beta$(6, 3)$ looks like



Nonparametric Bayesian Methods

Now we know what nonparametric and Bayesian mean.

What should we expect from nonparametric Bayesian methods?

- Complexity of our model should be allowed to grow as we get more data.

- Place a prior on an unbounded number of parameters.

Nonparametric Bayesian Methods

- *Parametric models* assume some finite set of parameters θ. Given the parameters, future predictions, x, are independent of the observed data, $\mathcal{D}$:

$$P(x|\theta, \mathcal{D}) = P(x|\theta)$$

  therefore θ capture everything there is to know about the data.

- So the complexity of the model is bounded even if the amount of data is unbounded. This makes them not very flexible.

Nonparametric Bayesian Methods

- *Non-parametric models* assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they can often be defined by assuming an *infinite dimensional* θ. Usually we think of θ as a *function*.

- The amount of information that θ can capture about the data $\mathcal{D}$ can grow as the amount of data grows. This makes them more flexible.

Bayesian nonparametrics

A simple framework for modelling complex data.

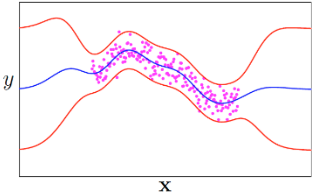
Nonparametric models can be viewed as having infinitely many parameters

Examples of non-parametric models:

Parametric	Non-parametric	Application
polynomial regression	Gaussian processes	function approx.
logistic regression	Gaussian process classifiers	classification
mixture models, k-means	Dirichlet process mixtures	clustering
hidden Markov models	infinite HMMs	time series
factor analysis / pPCA / PMF	infinite latent factor models	feature discovery
...		

Nonlinear regression and Gaussian processes

Consider the problem of nonlinear regression:
You want to learn a function f with error bars from data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$



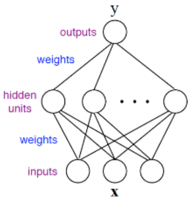
A Gaussian process defines a distribution over functions $p(f)$ which can be used for Bayesian regression:

$$p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})}$$

Let $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))$ be an n-dimensional vector of function values evaluated at n points $x_i \in \mathcal{X}$. Note, $\mathbf{f}$ is a random variable.

Definition: $p(f)$ is a Gaussian process if for *any* finite subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the marginal distribution over that subset $p(\mathbf{f})$ is multivariate Gaussian.

Neural networks and Gaussian processes



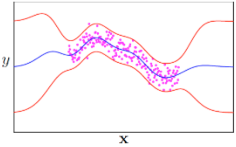
Bayesian neural network

Data: $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N} = (X, \mathbf{y})$
Parameters $\boldsymbol{\theta}$ are the weights of the neural net

parameter prior	$p(\boldsymbol{\theta}\|\boldsymbol{\alpha})$
parameter posterior	$p(\boldsymbol{\theta}\|\boldsymbol{\alpha}, \mathcal{D}) \propto p(\mathbf{y}\|X, \boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\alpha})$
prediction	$p(y'\|\mathcal{D}, \mathbf{x}', \boldsymbol{\alpha}) = \int p(y'\|\mathbf{x}', \boldsymbol{\theta})p(\boldsymbol{\theta}\|\mathcal{D}, \boldsymbol{\alpha})\, d\boldsymbol{\theta}$

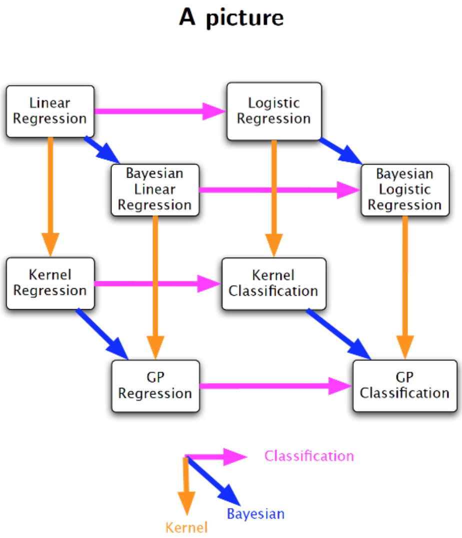
A **Gaussian process** models functions $y = f(\mathbf{x})$

A multilayer perceptron (neural network) with infinitely many hidden units and Gaussian priors on the weights $\rightarrow$ a GP (Neal, 1996)



See also recent work on Deep Gaussian Processes (Damianou and Lawrence, 2013)

Nonparametric Bayesian Methods



Next Lecture:

Popular NPB Models.