CPSC 540: Machine Learning

Non-Parametric Bayes

Mark Schmidt

University of British Columbia

Winter 2020

Stochastic Processes and Non-Parametric Bayes

- A stochastic process is an infinite collection of random variables $\{x^i\}$.
- Non-parametric Bayesian methods use priors defined on stochastic processes:
  - Allows extremely-flexible prior, and posterior complexity grows with data size.
  - Typically set up so that samples from posterior are finite-sized.
- The two most common priors are Gaussian processes and Dirichlet processes:
  - Gaussian processes define prior on space of functions (universal approximators).
  - Dirichlet processes define prior on space of probabilities (without fixing dimension).

Gaussian Processes

- Recall the partitioned form of a multivariate Gaussian

$$\mu = \begin{bmatrix}\mu_x, \mu_y\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix},$$

  and in this case the marginal $p(x)$ is a $\mathcal{N}(\mu_x, \Sigma_{xx})$ Gaussian.

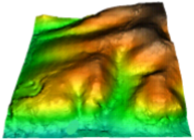
- Generalization of this to infinite set of variables is Gaussian processes (GPs):
  - Any finite set from collection follows a Gaussian distribution.

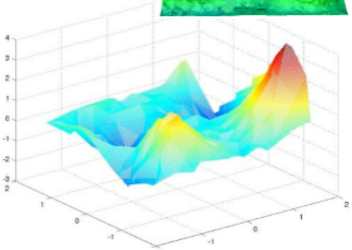
Gaussian Processes

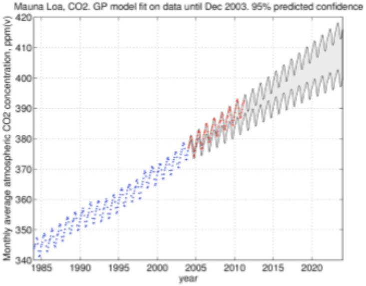
To date kriging has been used in a variety of disciplines, including the following:

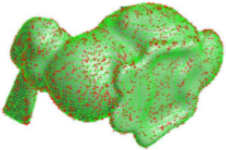
- Environmental science[5]
- Hydrogeology[6][7][8]
- Mining[9][10]
- Natural resources[11][12]
- Remote sensing[13]
- Real estate appraisal[14][15]

and many others.









Gaussian Processes

- GPs are specified by a mean function m and covariance function k,

$$m(x) = \mathbb{E}[f(x)], \quad k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))^T].$$

- Any finite sample $f(x)$ from a GP follows a $\mathcal{N}(m(x), k(x, x))$ distribution.
  - Analogous to partitioned Gaussian where $m(x) = \mu_x$ and $k(x, x) = \Sigma_{xx}$.

- We write that

$$f(x) \sim \mathsf{GP}(m(x), k(x, x')),$$

- As an example, we could have a zero-mean and linear covariance GP,

$$m(x) = 0, \quad k(x, x') = x^T x'.$$

Regression Models as Gaussian Processes

- As an example, predictions made by linear regression with Gaussian prior

$$f(x) = w^T \underbrace{\phi(x)}_{z}, \quad w \sim \mathcal{N}(0, \Sigma),$$

are a Gaussian process with mean function

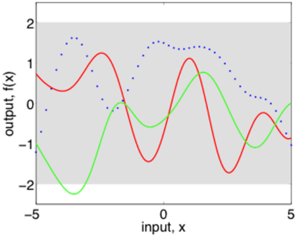
$$\mathbb{E}[f(x)] = \mathbb{E}[w^T \phi(x)] = \underbrace{\mathbb{E}[w]}_{0}^T \phi(x) = 0.$$

and covariance function

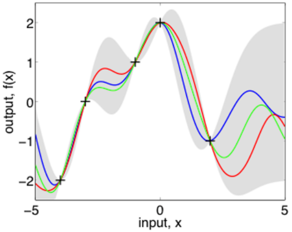
$$\mathbb{E}[f(x)f(x)^T] = \phi(x)^T \underbrace{\mathbb{E}[ww^T]}_{\Sigma} \phi(x') = \phi(x)\Sigma\phi(x') = k(x, x').$$

Gaussian Process Model Selection

- We can view a Gaussian process as a prior distribution over smooth functions.



(a), prior



(b), posterior

- Most common choice of covariance is Gaussian RBF.
  - Though "Matérn" kernel often works better.
- Is this related to using RBF kernels or the RBFs as the bases?
  - Yes, this is Bayesian linear regression plus the kernel trick.

Gaussian Process Model Selection

- So why do we care?
  - We can get estimate of uncertainty in the prediction.
  - We can use marginal likelihood to learn the kernel/covariance.
- Write kernel in terms of parameters, use empirical Bayes to learn kernel.
- Hierarchical approach: put a hyper-prior of types of kernels.
- Application: Bayesian optimization of non-convex functions:
  - Gradient descent is based on a Gaussian (quadratic) approximation of f.
  - Bayesian optimization is based on a Gaussian process approximation of f.
    - Can approximate non-convex functions.

Dirichlet Process

- Recall the basic mixture model:

$$p(x \mid \theta) = \sum_{c=1}^{k} \pi_c p(x \mid \theta_c).$$

- Non-parametric Bayesian methods allow us to consider infinite mixture model,

$$p(x \mid \theta) = \sum_{c=1}^{\infty} \pi_c p(x \mid \theta_c).$$

- Common choice for prior on π values is Dirichlet process:
  - Also called "Chinese restaurant process" and "stick-breaking process".
  - For finite datasets, only a fixed number of clusters have $\pi_c \neq 0$.
  - But don't need to pick number of clusters, grows with data size.

Dirichlet Process

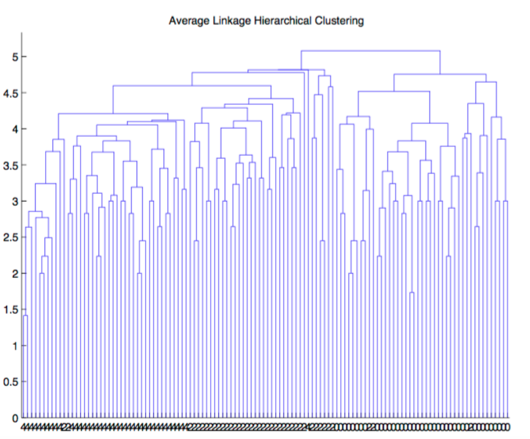
- Gibbs sampling in Dirichlet process mixture model in action:
  `https://www.youtube.com/watch?v=0Vh7qZY9sPs`

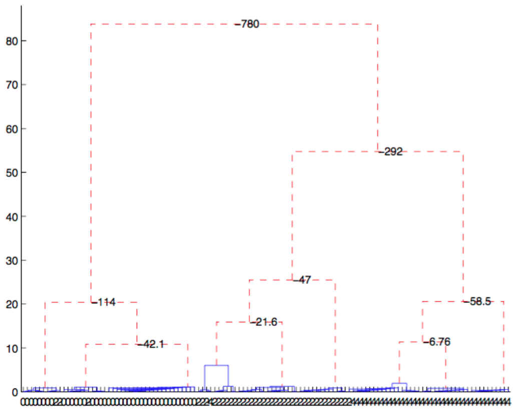
- We could alternately put a prior on k:
  - "Reversible-jump" MCMC can be used to sample from models of different sizes.
    - AKA "trans-dimensional" MCMC.

- There a variety of interesting variations on Dirichlet processes
  - Beta process ("Indian buffet process").
  - Hierarchical Dirichlet process.
  - Polya trees.
  - Infinite hidden Markov models.

Bayesian Hierarchical Clustering

- Hierarchical clustering of $\{0, 2, 4\}$ digits using classic and Bayesian method:

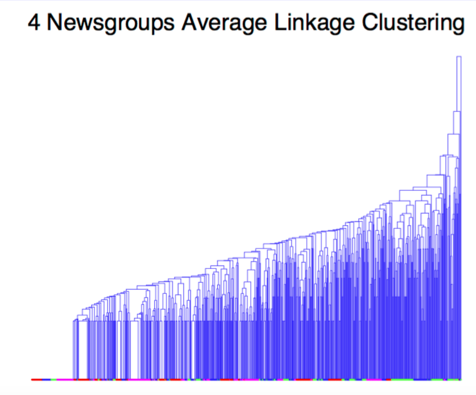


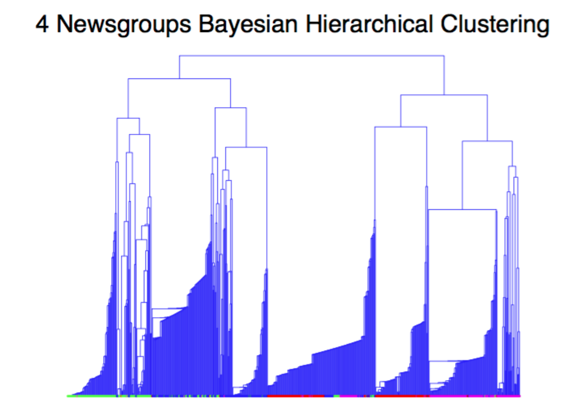


http://www2.stat.duke.edu/~kheller/bhcnew.pdf (y-axis represents distance between clusters)

Bayesian Hierarchical Clustering

- Hierarchical clustering of newgroups using classic and Bayesian method:





http://www2.stat.duke.edu/~kheller/bhcnew.pdf (y-axis represents distance between clusters)

Summary

- Non-Parametric Bayes use stochastic processes to model infinite spaces.
- Gaussian processes are priors over continuous functions.
- Dirichlet processes are priors over probability mass functions.
- Next time: new generative deep learning methods.