Nonparametric Bayesian Models, Dirichlet Process

Andrea Gesmundo

May, 15, 2009
Introduction

- Maximum Likelihood Estimation
- Bayesian Methods
- Beta Distribution
- Dirichlet Distribution
- Dirichlet Process
Problem:
Given the task of learning to predict a label $y$ for an observed variable $x$ based on an iid sample of training instances $x_1, \ldots, x_N$ and their labels $y_1, \ldots, y_N$
Problem:
Given the task of learning to predict a label $y$ for an observed variable $x$ based on an iid sample of training instances $x_1, ..., x_N$ and their labels $y_1, ..., y_N$

$x_n \in X, y_n \in Y$

Ex:
- $X$ is a set of words and $Y$ set of tags
Problem:
Given the task of learning to predict a label $y$ for an observed variable $x$ based on an iid sample of training instances $x_1, ..., x_N$ and their labels $y_1, ..., y_N$

$$x_n \in X, y_n \in Y$$

Ex:
- $X$ is a set of words and $Y$ set of tags
- Coin flip: $X$ contains single element, $Y = \{H, T\}$
Statistical Estimation Approach:
Devise a class of joint probability distributions (model) parameterized by $\theta$

$$p_\theta(x, y)$$
Statistical Estimation Approach:
Devise a class of joint probability distributions (model) parameterized by $\theta$

$$p_{\theta}(x, y)$$

Predicting the label for $x_n$ then becomes the task of finding

$$y_n = \arg\max_y p_{\theta}(y, x_n) = \arg\max_y p_{\theta}(y | x_n)$$
Maximum-Likelihood Estimation

Common method for determining the parameters $\theta$. MLE choose the $\theta$ that maximizes the probability of the training corpus

$D = (x_1, y_1) \cdots (x_N, y_N)$

$$\hat{\theta} = \arg \max_{\theta} p_{\theta}(D) = \arg \max_{\theta} p_{\theta}(x_1, y_1) \cdots p_{\theta}(x_N, y_N)$$
Maximum-Likelihood Estimation

Common method for determining the parameters $\theta$. MLE choose the $\theta$ that maximizes the probability of the training corpus $D = (x_1, y_1) \ldots (x_N, y_N)$

$$\hat{\theta} = \arg \max \theta \ p_\theta(D) = \arg \max \theta \ p_\theta(x_1, y_1) \cdots p_\theta(x_N, y_N)$$

MLE is prone to overfitting: The estimated distribution only acknowledges events that occurred in the training data.
$x$ is a constant, $y \in \{H, T\}$ and $\theta \equiv \{\theta_1, \theta_2\}$

Model as Bernoulli distribution:

$$p_\theta(x, y) = \begin{cases} 
\theta_1 & \text{if } y=H \\
\theta_2 & \text{if } y=T 
\end{cases}$$

Where $\theta_2 = 1 - \theta_1$. 
\( x \) is a constant, \( y \in \{H, T\} \) and \( \theta \equiv \{\theta_1, \theta_2\} \)

Model as Bernoulli distribution:

\[
p_\theta(x, y) = \begin{cases} 
\theta_1 & \text{if } y = H \\
\theta_2 & \text{if } y = T 
\end{cases}
\]

Where \( \theta_2 = 1 - \theta_1 \).

Therefore the likelihood function \( p_\theta(D) \) is the Binomial:

\[
p_\theta(D) = \theta_1^H \theta_2^T
\]
Find the parameters with MLE

$$\hat{\theta} = \max \arg_{\theta} p_{\theta}(D) = \left\{ \frac{\#H}{\#H + \#T}, \frac{\#T}{\#H + \#T} \right\}$$

Where $\#H$ is the number of heads in the training data $D$. 
Find the parameters with MLE

\[ \hat{\theta} = \max \arg_y \ p_\theta(D) = \left\{ \frac{\# H}{\# H + \# T}, \frac{\# T}{\# H + \# T} \right\} \]

Where \#_H is the number of heads in the training data D.

Imagine what happens when we toss the sequence \{H, T, T\}

\[ \hat{\theta} = \left\{ \frac{1}{3}, \frac{2}{3} \right\} \]

Our estimator would be biased toward tails
How is possible to avoid overfitting?
How is possible to avoid overfitting?

Bayesian Model:
- Treat the model parameters as random variables
- Assign a prior distribution to the parameter

prior: $p(\theta)$, is a distribution over distributions
Graphical representation:
All the possible value of $\theta$ can be represented with a segment
Graphical representation:
Associate a probability density value to each point = parameter vector = distribution
Bayesian models work with a posterior distribution:

\[ p(\theta|D) = \frac{p(D|\theta) \ p(\theta)}{p(D)} \]
Bayesian models work with a posterior distribution:

\[ p(\theta|D) = \frac{p(D|\theta) p(\theta)}{p(D)} \propto p(D|\theta) p(\theta) \]
Bayesian models works with a posterior distribution:

$$p(\theta|D) = \frac{p(D|\theta) \ p(\theta)}{p(D)} \propto p(D|\theta) \ p(\theta)$$

Bayes formula comes from:

$$p(D|\theta) \ p(\theta) = p(D, \theta) = p(\theta, D) = p(\theta|D) \ p(D)$$
How do we work with a $p(\theta|D)$ instead a $\hat{\theta}$?
Bayesian models

How do we work with a $p(\theta|D)$ instead a $\hat{\theta}$?

MLE: $y_n = \arg \max_y p_{\hat{\theta}}(y|x_n)$
How do we work with a $p(\theta|D)$ instead a $\hat{\theta}$?

**MLE** : $y_n = \arg \max_y p_{\hat{\theta}}(y|x_n)$

**Bayesian: Marginalization**

$$P(y_n|D) = \int_\theta P(y_n|\theta) \ p(\theta|D) \ d\theta$$
We can compute Likelihood from Data (Binomial):

\[ p(D|\theta) = \theta_1^{\#H} \theta_2^{\#T} \]

Choose prior:

- It is convenient to chose a prior that gives a posterior that is easy to integrate
We can compute Likelihood from Data (Binomial):

\[ p(D|\theta) = \theta_1^{\#H} \theta_2^{\#T} \]

Choose prior:

- It is convenient to chose a prior that gives a posterior that is easy to integrate
- Given the likelihood we can chose the prior in such a way the resulting posterior distribution is of the same form of the prior and thus has a closed-form solution
We can compute Likelihood from Data (Binomial):

\[ p(D|\theta) = \theta^H_1 \theta^T_2 \]

Choose prior:

- It is convenient to choose a prior that gives a posterior that is easy to integrate
- Given the likelihood we can choose the prior in such a way that the resulting posterior distribution is of the same form as the prior and thus has a closed-form solution
- Such priors are called *Conjugate*
We can compute Likelihood from Data (Binomial):

\[ p(D|\theta) = \theta_1^{#H} \theta_2^{#T} \]

Choose prior:
The prior of the Binomial is the Beta distribution: *Beta*(\(\alpha_H, \alpha_T\)) given by the pdf:

\[ p(\theta; \alpha_H, \alpha_T) = \frac{\theta_1^{(\alpha_H-1)} \theta_2^{(\alpha_T-1)}}{C} \]
Posterior:

\[
p(\theta|D) \propto p(D|\theta) \ p(\theta) \\
\propto \theta_1^{\#H} \ \theta_2^{\#T} \ \theta_1^{(\alpha_H-1)} \ \theta_2^{(\alpha_T-1)} \\
= \theta_1^{(\#H+\alpha_H-1)} \ \theta_2^{(\#T+\alpha_T-1)}
\]

Therefore we have:

\[
\theta|D \sim Beta(\#_H + \alpha_H, \#_T + \alpha_T)
\]
Estimate probability of a 'head' toss

\[
P(H|D) = \int_\theta P(H|\theta) \, p(\theta|D) \, d\theta
\]

\[
= \int_\theta \theta_1 \, p(\theta|D) \, d\theta
\]

\[
= E[\theta_1|D]
\]

\[
= \frac{\#H + \alpha_H}{\#H + \alpha_H + \#T + \alpha_T}
\] (2)
Coin Flip Example with Bayesian model

Estimate probability of a ’head’ toss

\[
P(H|D) = \int_{\theta} P(H|\theta) \ p(\theta|D) \ d\theta
\]

\[
= \int_{\theta} \theta_1 \ p(\theta|D) \ d\theta
\]

\[
= E[\theta_1|D]
\]

\[
= \frac{\#_H + \alpha_H}{\#_H + \alpha_H + \#_T + \alpha_T}
\]

- When sample D is small \(\alpha\) parameters dominate
Coin Flip Example with Bayesian model

Estimate probability of a 'head' toss

\[ P(H|D) = \int_{\theta} P(H|\theta) \ p(\theta|D) \ d\theta \]

\[ = \int_{\theta} \theta_1 \ p(\theta|D) \ d\theta \]

\[ = E[\theta_1|D] \]

\[ = \frac{\#H + \alpha_H}{\#H + \alpha_H + \#T + \alpha_T} \]

- When sample D is small \( \alpha \) parameters dominate
- MLE interpretation: prepend D with \( \alpha_H \) head tosses and \( \alpha_T \) tail tosses

Andrea Gesmundo
Beta Distribution Examples

\[
\begin{align*}
\alpha = \beta &= 0.5 \\
\alpha = 5, \beta &= 1 \\
\alpha = 1, \beta &= 3 \\
\alpha = 2, \beta &= 2 \\
\alpha = 2, \beta &= 5
\end{align*}
\]
Rewrite Beta distribution parameters $\alpha_1, \alpha_2$ as:

- **Concentration Parameter, $\alpha$:**
  \[ \alpha = \alpha_1 + \alpha_2 \]

- **Distribution Mean, $(\alpha'_1, \alpha'_2)$:**
  \[ \alpha'_i = \frac{\alpha_i}{\alpha} \]

If $\alpha > 2$ distribution is concentrated near the mean
If $\alpha < 2$ distribution is concentrated far from the mean, sparsity
Generalize for all ‘N’

\[ Y = \{ H, T \} \]

\[ \theta = \{ \theta_1, \theta_2 \} \quad \text{with } \theta_1 + \theta_2 = 1 \]

\[ p(D|\theta) = \theta_1^{\#_H} \theta_2^{\#_T} \]

\[ p(\theta; \alpha_H, \alpha_T) \propto \theta_1^{(\alpha_H-1)} \theta_2^{(\alpha_T-1)} \]

\[ p(\theta|D) \propto p(D|\theta) \; p(\theta) \propto \theta_1^{(\#_H+\alpha_H-1)} \theta_2^{(\#_T+\alpha_T-1)} \]

\[ P(H|\theta) = \theta_1 \]

\[ P(H|D) = \frac{\#_H + \alpha_H}{\#_H + \alpha_H + \#_T + \alpha_T} \]

Andrea Gesmundo
Generalize for all 'N'

\[ Y = \{ \psi_1, \psi_2 \} \]
\[ \theta = \{ \theta_1, \theta_2 \} \quad \text{with} \quad \theta_1 + \theta_2 = 1 \]
\[ p(D|\theta) = \theta_1^{\#\psi_1} \theta_2^{\#\psi_2} \]
\[ p(\theta; \alpha_1, \alpha_2) \propto \theta_1^{(\alpha_1-1)} \theta_2^{(\alpha_2-1)} \]
\[ p(\theta|D) \propto p(D|\theta) \ p(\theta) \propto \theta_1^{(\#\psi_1+\alpha_2-1)} \theta_2^{(\#\psi_2+\alpha_2-1)} \]
\[ P(\psi_i|\theta) = \theta_i \]
\[ P(\psi_i|D) = \frac{\#\psi_i + \alpha_i}{\#\psi_1 + \#\psi_2 + \alpha_1 + \alpha_2} \]
Generalize for all ’N’

\[ Y = \{\psi_1, ..., \psi_N\} \]

\[ \theta = \{\theta_1, ..., \theta_N\} \quad \text{with} \quad \sum_{n=1}^{N} \theta_n = 1 \]

\[ p(D|\theta) = \prod_{n=1}^{N} \theta_n^{\#\psi_n} \]

\[ p(\theta; \alpha_1, ..., \alpha_N) \propto \prod_{n=1}^{N} \theta_n^{(\alpha_n-1)} \]

\[ p(\theta|D) \propto p(D|\theta) \ p(\theta) \propto \prod_{n=1}^{N} \theta_n^{(\#\psi_n+\alpha_n-1)} \]

\[ P(\psi_i|\theta) = \theta_i \]

\[ P(\psi_i|D) = \frac{\#\psi_i + \alpha_i}{\sum_{n=1}^{N} \#\psi_n + \sum_{n=1}^{N} \alpha_n} \]
Dirichlet distribution is a multi-parameter generalization of the Beta distribution.

It is the Conjugate of the Multinomial.

It defines a distribution over distribution.

\[
p(\theta; \alpha_1, \ldots, \alpha_N) = \frac{\prod_{n=1}^{N} \theta_n^{(\alpha_n-1)}}{C}
\]

\[
\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_N)
\]
Dirichlet Distribution with N=3

Dirichlet(1,1,1)  Dirichlet(2,2,2)  Dirichlet(10,10,10)

Dirichlet(2,10,2)  Dirichlet(2,2,10)  Dirichlet(0.9,0.9,0.9)
Rewrite Dirichlet distribution parameters $\alpha_1, \ldots, \alpha_N$ as:

- **Concentration Parameter, $\alpha$:**
  $$\alpha = \sum_{n=1}^{N} \alpha_n$$

- **Distribution Mean, $(\alpha'_1, \ldots, \alpha'_N)$:**
  $$\alpha'_i = \frac{\alpha_i}{\alpha}$$

If $\alpha > N$ distribution is concentrated near the mean
If $\alpha < N$ distribution is concentrated far from the mean, sparsity
Note that \((\alpha'_1, \ldots, \alpha'_N) = h\) can be considered a distribution on the event space \(Y\) as \(\theta = (\theta_1, \ldots, \theta_N)\)
• Note that \((\alpha'_1, ..., \alpha'_N) = h\) can be considered a distribution on the event space \(Y\) as \(\theta = (\theta_1, ..., \theta_N)\).

• With \(\alpha \to \infty\) we have \(\theta \to h = (\alpha'_1, ..., \alpha'_N)\).
Note that \((\alpha'_1, \ldots, \alpha'_N) = h\) can be considered a distribution on the event space \(Y\) as \(\theta = (\theta_1, \ldots, \theta_N)\).

With \(\alpha \to \infty\) we have \(\theta \to h = (\alpha'_1, \ldots, \alpha'_N)\).

Note that:

\[
\text{Dirichlet}(\alpha_1, \ldots, \alpha_N) \equiv \text{Dirichlet}(\alpha \alpha'_1, \ldots, \alpha \alpha'_N)
\]

or \(\text{Dirichlet}(\alpha, h)\)
Polya Urn Model

- Many probability distributions can be obtained using urn models

- Polya Urn Model is the one that corresponds to the Dirichlet Distribution
Consider an urn with $\alpha$ balls
Consider an urn with $\alpha$ balls

Each ball can have one of $N$ different colours
Consider an urn with $\alpha$ balls

Each ball can have one of $N$ different colours

There are $\alpha_n$ balls of colour 'n', with $1 \leq n \leq N$
Polya Urn Model

Consider an urn with $\alpha$ balls

Each ball can have one of $N$ different colours

There are $\alpha_n$ balls of colour ’n’, with $1 \leq n \leq N$

We draw balls at random, repleace the ball we drew with two balls of the same colour.
Polya Urn Model

\[ P(ball_1 = n) = \frac{\alpha n}{\alpha} \]
Polya Urn Model

\[ P(ball_1 = n) = \frac{\alpha_n}{\alpha} \]

\[ P(ball_2 = n) = \frac{\#_{n,1} + \alpha_n}{1 + \alpha} \]

\[ \vdots \]

\[ P(ball_{k+1} = n) = \frac{\#_{n,k} + \alpha_n}{k + \alpha} \]
Polya Urn Model

\[ P(ball_1 = n) = \frac{\alpha_n}{\alpha} \]

\[ P(ball_2 = n) = \frac{\#_{n,1} + \alpha_n}{1 + \alpha} \]

\[ \vdots \]

\[ P(ball_{k+1} = n) = \frac{\#_{n,k} + \alpha_n}{k + \alpha} \]

with \( k \rightarrow \infty \) the proportions of different colors in the urn will be distributed according to the \( Dirichlet(\alpha_1, \ldots, \alpha_N) \)
Generalize for all 'N'

\[ Y = \{ \psi_1, ..., \psi_N \} \]

\[ \theta = \{ \theta_1, ..., \theta_N \} \quad \text{with} \quad \sum_{n=1}^{N} \theta_n = 1 \]

\[ p(D|\theta) = \prod_{n=1}^{N} \theta_n^{\#\psi_n} \]

\[ p(\theta; \alpha_1, ..., \alpha_N) \propto \prod_{n=1}^{N} \theta_n^{(\alpha_n-1)} \]

\[ p(\theta|D) \propto p(D|\theta) \cdot p(\theta) \propto \prod_{n=1}^{N} \theta_n^{(\#\psi_n+\alpha_n-1)} \]

\[ P(\psi_i|\theta) = \theta_i \]

\[ P(\psi_i|D) = \frac{\#\psi_i + \alpha_i}{\sum_{n=1}^{N} \#\psi_n + \sum_{n=1}^{N} \alpha_n} \]
Generalize for $N \rightarrow \infty$

- $Y$ becomes a continuous space
Generalize for $N \to \infty$

- $Y$ becomes a continuous space
- $\theta$ was a PMF and becomes a PDF
Generalize for $N \to \infty$

- $Y$ becomes a continuous space

- $\theta$ was a PMF and becomes a PDF

- In a PDF we are interested in probabilities of $y \in \Psi_i$ with $\Psi_i \subseteq Y$
Generalize for $N \to \infty$

- $Y$ becomes a continuous space

- $\theta$ was a PMF and becomes a PDF

- In a PDF we are interested in probabilities of $y \in \Psi_i$ with $\Psi_i \subseteq Y$

- $\theta_i$ was the $P(y = \psi_i)$, now we write $\theta(\Psi_i)$ for $P(y \in \Psi_i)$
Dirichelt Process

DP is an extension of the Dirichelet distribution to continuous spaces
DP is an extension of the Dirichelet distribution to continuous spaces

Definition:
Let $H$ be a distribution over event space $Y$, and let $\alpha$ be a positive real number.
A distribution $\theta$ over $Y$ is said to be drawn from a Dirichlet Process with base distribution $H$ and concentration parameter $\alpha$

$$\theta \sim DP(\alpha, H)$$

if for every measurable partition $\Psi_1, \ldots, \Psi_r$ of $Y$

$$(\theta(\Psi_1), \ldots, \theta(\Psi_r)) \sim Dirichlet(\alpha H(\Psi_1), \ldots, \alpha H(\Psi_r))$$
Dirichlet Process Properties

- $E[\theta(\psi)] = H(\psi)$
- $V[\theta(\psi)] = \frac{H(\psi)(1 - H(\psi))}{\alpha + 1}$
Dirichlet Process Properties

- $E[\theta(\psi)] = H(\psi)$

- $V[\theta(\psi)] = \frac{H(\psi)(1 - H(\psi))}{\alpha + 1}$

Note:
From the discrete case:

- If $\alpha > N$ distribution is concentrated near the mean
- If $\alpha < N$ distribution is concentrated far from the mean, sparsity

In the continuous case $N \to \infty$ and so $\alpha \ll N$
So we can imagine that the draw from the DP will be really sparse
This observation introduce to the most important DP Property
It may seem that $\theta$ is continuous since $H$ is continuous.
• It may seem that $\theta$ is continuous since $H$ is continuous

• But the draw from the DP are so sparse that $\theta$ consist of countably infinite point probability masses
It may seem that $\theta$ is continuous since $H$ is continuous.

But the draw from the DP are so sparse that $\theta$ consist of countably infinite point probability masses.

Therefore values observed from a DP previously have a non-zero probability of occurring again.
The Urn is empty at the beginning
- The Urn is empty at the beginning
- $Y$ is a continuous interval of colours
The Urn is empty at the beginning

\( Y \) is a continuous interval of colours

We select an \( \alpha > 0 \) and a distribution \( H \) over \( Y \)
The Urn is empty at the beginning

\( Y \) is a continuous interval of colours

We select an \( \alpha > 0 \) and a distribution \( H \) over \( Y \)

In each subsequent step \( k + 1 \) either:

- a colour \( \psi_{k+1} \) is drawn from \( H \) with probability \( \frac{\alpha}{\alpha+k} \), and a ball is colored with it and added to the urn

- or with probability \( \frac{k}{\alpha+k} \) a ball is drawn from the urn, its color is used to color a new ball and both balls are added to the urn.
Blackwell-MacQueen formula

\[ P(y_{k+1} = \psi_n | y_{1:k}) = \begin{cases} \frac{\#\psi_n, k}{\alpha + k} & \text{if } \exists j \leq k, \ s.t. \ y_j = \psi_n \\ \frac{\alpha}{\alpha + k} H(\psi_n) & y_j \neq \psi_n, \ \forall \ 1 \leq j \leq k \end{cases} \]
A way to construct a distribution $G \sim DP(\alpha, H)$

$$\beta_k \sim Beta(1, \alpha)$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

$$\psi_k \sim H$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\psi_k}$$

Recursively breaking the remaining stick ar ratio $\beta_k$ to obtain $\pi_k$

And assign a proportional probability to the value $\psi_k \in Y$ extracted from $H$
Combining entries of probability vectors preserves Dirichlet property, for example:

\[(\pi_1, \pi_2, \ldots, \pi_n) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_n)\]

\[\Rightarrow (\pi_1 + \pi_2, \ldots, \pi_n) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \ldots, \alpha_n)\]
Combining entries of probability vectors preserves Dirichlet property, for example:

\[(\pi_1, \pi_2, ..., \pi_n) \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_n)\]

\[\Rightarrow (\pi_1 + \pi_2, ..., \pi_n) \sim Dirichlet(\alpha_1 + \alpha_2, ..., \alpha_n)\]

Generally, if \((l_1, ..., l_k)\) is a partition of \((1, ..., n)\).

\[(\sum_{i \in l_1} \pi_1, ..., \sum_{i \in l_k} \pi_n) \sim Dirichlet(\sum_{i \in l_1} \alpha_1, ..., \sum_{i \in l_k} \alpha_i)\]
Splitting entries of probability vectors preserves Dirichlet property, for example:

\[
\begin{align*}
(p_1, p_2, \ldots, p_n) & \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_n) \\
(q_1, q_2) & \sim \text{Dirichlet}(\alpha_1 \beta_1, \alpha_2 \beta_2)
\end{align*}
\]

with \(\beta_1 + \beta_2 = 1\)
Splitting entries of probability vectors preserves Dirichlet property, for example:

\[(\pi_1, \pi_2, \ldots, \pi_n) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_n)\]

\[(\tau_1, \tau_2) \sim \text{Dirichlet}(\alpha_1 \beta_1, \alpha_2 \beta_2)\]

with \(\beta_1 + \beta_2 = 1\)

\[\Rightarrow \quad (\pi_1 \tau_1, \pi_1 \tau_2, \pi_2, \ldots, \pi_n) \sim \text{Dirichlet}(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2, \ldots, \alpha_n)\]
A Dirichlet Process is an "infinitely decimated" Dirichlet distribution:

\[ \begin{align*}
1 & \sim \text{Dirichlet}(\alpha) \\
(\pi_1, \pi_2) & \sim \text{Dirichlet}(\alpha/2, \alpha/2) \quad \pi_1 + \pi_2 = 1 \\
(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) & \sim \text{Dirichlet}(\alpha/4, \alpha/4, \alpha/4, \alpha/4) \quad \pi_{i1} + \pi_{i2} = \pi_i \\
& \vdots \\
\text{Each decimation step involves drawing from a Beta Distribution,} \\
\text{and multiplying into the relevant entry.}
\end{align*} \]
Dirichlet Processes are used for:

- Density/Function Estimation
- Semiparametric Modelling
- Sidestepping model averaging/selection
- Parametric Function estimation (e.g. regression, classification)
  
  Data: $x = \{x_1, x_2, \ldots\}$, $y = \{y_1, y_2, \ldots\}$
  
  Model: $y_i = f(x_i \mid w) + \mathcal{N}(0, \sigma^2)$
Function Estimation

- Parametric Function estimation (e.g. regression, classification)
  Data: \( x = \{x_1, x_2, \ldots \} \), \( y = \{y_1, y_2, \ldots \} \)
  Model: \( y_i = f(x_i|w) + \mathcal{N}(0, \sigma^2) \)
- Prior over parameters
  \( p(w) \)
Parametric Function estimation (e.g. regression, classification)

Data: $x = \{x_1, x_2, \ldots\}$, $y = \{y_1, y_2, \ldots\}$

Model: $y_i = f(x_i|w) + \mathcal{N}(0, \sigma^2)$

Prior over parameters

$$p(w)$$

Prosterior over parameters

$$p(w|x, y) = \frac{p(w)p(y|x, w)}{p(y|x)}$$
Parametric Function estimation (e.g. regression, classification)

Data: \( x = \{x_1, x_2, \ldots \} \), \( y = \{y_1, y_2, \ldots \} \)

Model: \( y_i = f(x_i|w) + \mathcal{N}(0, \sigma^2) \)

Prior over parameters

\( p(w) \)

Posterior over parameters

\[
 p(w|x, y) = \frac{p(w)p(y|x, w)}{p(y|x)}
\]

Prediction with posteriors

\[
 p(y_i|x_i, x, y) = \int p(y_i|x_i, w)p(w|x, y)dw
\]
Parametric Function estimation (e.g. regression, classification)
Data: \( x = \{x_1, x_2, \ldots \} \), \( y = \{y_1, y_2, \ldots \} \)
Model: \( y_i = f(x_i) + \mathcal{N}(0, \sigma^2) \)

Prior over functions
\[ f \sim GP(\alpha, H) \]

Posterior over functions
\[ p(f|x, y) = \frac{p(f)p(y|x, f)}{p(y|x)} \]

Prediction with posteriors
\[ p(y_i|x_i, x, y) = \int p(y_i|x_i, f)p(f|x, y)df \]
Parametric density estimation (e.g. mixture models)

Data: \( x = \{x_1, x_2, \ldots \} \)

Model: \( x_i|w \sim F(\cdot|w) \)

Prior over parameters

\[ p(w) \]

Posterior over parameters

\[ p(w|x) = \frac{p(w)p(x|w)}{p(x)} \]

Prediction with posteriors

\[ p(x_i|x) = \int p(x_i|w)p(w|x)dw \]
Bayesian nonparametric density estimation with Dirichlet processes

Data: $x = \{x_1, x_2, \ldots\}$
Model: $x_i \sim F$

Prior over distribution

$$F \sim DP(\alpha, H)$$

Posterior over distributions

$$p(F|x) = \frac{p(F)p(x|F)}{p(x)}$$

Prediction with posteriors

$$p(x_i|x) = \int p(x_i|F)p(F|x)dF$$
Bayesian nonparametric density estimation with Dirichlet processes

Data: \( x = \{x_1, x_2, \ldots \} \)

Model: \( x_i \sim F \)

Prior over distribution
\[
F \sim DP(\alpha, H)
\]

Posterior over distributions
\[
p(F|x) = \frac{p(F)p(x|F)}{p(x)}
\]

Prediction with posteriors
\[
p(x_i|x) = \int p(x_i|F)p(F|x)dF = \int F'(x_i)p(F|x)dF
\]
Parametric density estimation (e.g. mixture models)

Data: \( x = \{x_1, x_2, \ldots \} \)

Models: \( p(\theta_k | M_k), p(x | \theta_k, M_k) \)

Marginal likelihood

\[
p(x | M_k) = \int p(x | \theta_k, M_k) p(\theta_k | M_k) d\theta_k
\]

Model selection

\[
M = \arg\max_{M_k} p(x | M_k)
\]

Model averaging

\[
p(x_i | x) = \sum_{M_k} p(x_i | M_k) p(M_k | x)
\]
Model Selection/Averaging

- Parametric density estimation (e.g. mixture models)
  Data: \( x = \{x_1, x_2, \ldots \} \)
  Models: \( p(\theta_k|\mathcal{M}_k), p(x|\theta_k, \mathcal{M}_k) \)

- Marginal likelihood
  \[
  p(x|\mathcal{M}_k) = \int p(x|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k
  \]

- Model selection
  \[
  \mathcal{M} = \arg\max_{\mathcal{M}_k} p(x|\mathcal{M}_k)
  \]

- Model averaging
  \[
  p(x_i|x) = \sum_{\mathcal{M}_k} p(x_i|\mathcal{M}_k)p(\mathcal{M}_k|x)
  \]
  \[
  = \sum_{\mathcal{M}_k} p(x_i|\mathcal{M}_k)\frac{p(x|\mathcal{M}_k)p(\mathcal{M}_k)}{p(x)}
  \]
Marginal likelihood is usually extremely hard to compute.

\[ p(x|M_k) = \int p(x|\theta_k, M_k)p(\theta_k|M_k)d\theta_k \]

Model selection/averaging is to prevent underfitting and overfitting

But reasonable and proper Bayesian methods should not overfit

Use a really large model \( M_\infty \), and let the data speak for themselves.
A finite mixture model is defined as follows:

\[ \phi_k \sim H \]
\[ \pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \]
\[ z_i | \pi \sim \text{Discrete}(\pi) \]
\[ x_i | \phi_{z_i} \sim F(\cdot | \phi_{z_i}) \]

- Model selection/averaging over:
  - Hyperparameters in \( H \).
  - Dirichlet parameter \( \alpha \).
  - Number of components \( K \).

- Determining \( K \) hardest.
Imagine that \( K \gg 0 \) is really large.

If parameters \( \phi_k \) and mixing proportions \( \pi \) integrated out, the number of latent variables left does not grow with \( K \)—no overfitting.

At most \( n \) components will be associated with data, aka “active”.

Usually, the number of active components is much less than \( n \).

This gives an infinite mixture model.