Bayesian Methods (Lab)

David Rosenberg

New York University

November 1, 2015

Coin Flipping

- **Parameter space** $\theta \in \Theta = [0,1]$:

$$\mathbb{P}(\text{Heads} \mid \theta) = \theta.$$

- Data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$
  - n_h: number of heads
  - n_t: number of tails
- **Likelihood model** (Bernoulli Distribution):

$$p(\mathcal{D} \mid \theta) = \theta^{n_h} (1-\theta)^{n_t}$$

  - (probability of getting the flips in the order they were received)

Coin Flipping: Beta Prior

- **Prior:**

$$
\begin{aligned}
\theta &\sim \mathrm{Beta}(\alpha, \beta) \\
p(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}
\end{aligned}
$$

Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons `http://commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg`.

Coin Flipping: Beta Prior

- **Prior:**

$$\begin{aligned} \theta &\sim \text{Beta}(h, t) \\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1} \end{aligned}$$

- **Mean of Beta distribution:**

$$\mathbb{E}\theta = \frac{h}{h+t}$$

Coin Flipping: Posterior

- **Prior:**

$$\begin{aligned} \theta &\sim \text{Beta}(h,t) \\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1} \end{aligned}$$

- **Likelihood model**:

$$p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- **Posterior density:**

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto p(\theta)p(\mathcal{D} \mid \theta) \\ &\propto \theta^{h-1}(1-\theta)^{t-1} \times \theta^{n_h}(1-\theta)^{n_t} \\ &= \theta^{h-1+n_h}(1-\theta)^{t-1+n_t} \end{aligned}$$

Posterior is Beta

- **Prior:**

$$\begin{aligned}\theta &\sim \mathrm{Beta}(h,t)\\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}\end{aligned}$$

- **Posterior density:**

$$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$$

- **Posterior is in the beta family**:

$$\theta \mid \mathcal{D} \sim \mathrm{Beta}(h+n_h, t+n_t)$$

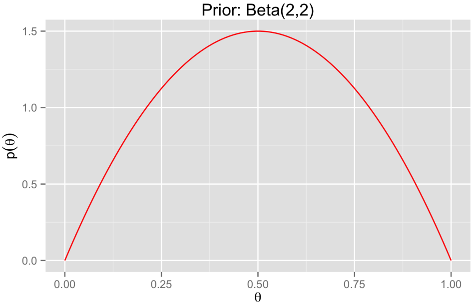
- **Interpretation**:
  - Prior initializes our counts with h heads and t tails.
  - Posterior increments counts by observed n_h and n_t.

Example: Coin Flipping

- Suppose we have a coin, possibly biased

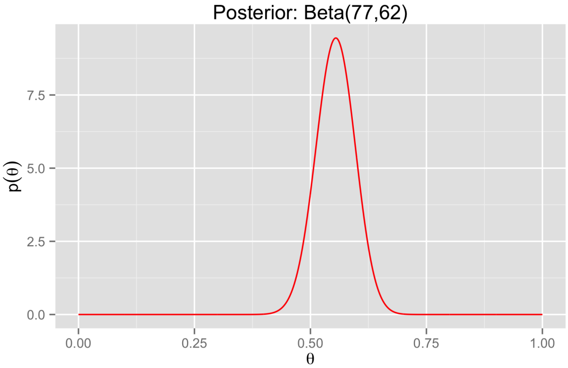
$$\mathbb{P}(\text{Heads} \mid \theta) = \theta.$$

- **Parameter space** $\theta \in \Theta = [0, 1]$.
- **Prior distribution:** $\theta \sim \text{Beta}(2, 2)$.



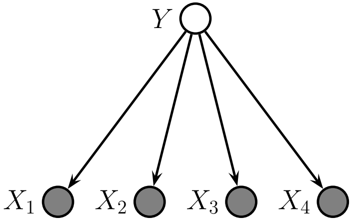
Example: Coin Flipping

- Next, we gather some data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$:
- Heads: 75 Tails: 60
  - $\hat{\theta}_{\text{MLE}} = \frac{75}{75+60} \approx 0.556$
- **Posterior distribution:** $\theta \mid \mathcal{D} \sim \textbf{Beta}(77, 62)$:



Naive Bayes: A Generative Model for Classification

- $\mathcal{X} = \left\{ \left(X_1, X_2, X_3, X_4 \right) \in \{0,1\}^4 \right\}$ $\mathcal{Y} = \{0,1\}$ be a class label.
- Consider the Bayesian network depicted below:



- BN structure implies joint distribution factors as:

$$p(x_1, x_2, x_3, x_4, y) = p(y)p(x_1 \mid y)p(x_2 \mid y)p(x_3 \mid y)p(x_4 \mid y)$$

- Features $X_1, \ldots, X_4$ are independent given the class label Y.

KPM Figure 10.2(a).

Example: Message Classification

- $\mathcal{X} = \{\text{Message Text}\}$
- $\mathcal{Y} = \{\text{BUSINESS}, \text{PERSONAL}\}$
- **Training Data**
  - BUSINESS
    - "Lunch meeting?"
    - "Expenses submitted EOM."
    - "LOL"
  - PERSONAL
    - "Meet for lunch? EOM"
    - "LOL"

Bag of Words Representation (Bernoulli Version)

- Represent a message by the set of words it contains:
  - ignores word order
  - ignores word count (some bag of words models keep the count)
  - typically ignores punctuation and capitalization
- Generate vocabulary from training data:

$$W = \{\text{eom},\text{expenses},\text{for},\text{lol},\text{lunch},\text{meet},\text{meeting},\text{submitted},\text{UNKNOWN}\}$$

  - Add in an UNKNOWN value, in case we encounter new words in deployment.
- Message M is represented by binary vector of length $|W| = 9$.

Bag of Words Representation (Bernoulli Version)

- Input: "Lunch? EOM" $\implies$ $M = \{\text{lunch}, \text{eom}\}$:
- Vector representation: $x = \left(x_1 \ldots, x_{|W|}\right)$

Word (w)	x_w
lunch	1
meeting	0
expenses	0
submitted	0
eom	1
meet	0
for	0
lol	0
UNKNOWN	0

Bernoulli Naive Bayes Model

- Joint probability of message $x = \left(x_1, \ldots, x_{|W|}\right)$ and class y is

$$p(x,y) = p(y)\prod_{i=1}^{|W|} p(x_i \mid y),$$

where each $x_i \in \{0,1\}$, and $y \in \{\mathrm{B}, \mathrm{P}\}$.

- We need to estimate:

$$\begin{gathered}\mathbb{P}(Y = \mathrm{B}) \\ \mathbb{P}(Y = \mathrm{P}) \\ \mathbb{P}(X_w = 1 \mid Y = \mathrm{B}) \; \forall w \in W \\ \mathbb{P}(X_w = 1 \mid Y = \mathrm{P}) \; \forall w \in W\end{gathered}$$

Bernoulli Naive Bayes: Parameter Estimation

- Using relative frequencies in training, we have:

$$\hat{p}(Y = B) = 3/5 \quad \hat{p}(Y = P) = 2/5$$

and

Word (w)	$\hat{p}(X_w = 1 \mid \mathrm{B})$	$\hat{p}(X_w = 1 \mid \mathrm{P})$
lunch	1/3	1/2
meeting	1/3	0
expenses	1/3	0
submitted	1/3	0
eom	1/3	1/2
meet	0	1/2
for	0	1/2
lol	1/3	1/2
UNKNOWN	0	0

Naive Bayes Prediction for "Lunch? EOM"

Word (w)	x_w	$\hat{p}(X_w = 1 \mid \text{B})$	$\hat{p}(x_w \mid B)$	$\hat{p}(X_w = 1 \mid \text{P})$	$\hat{p}(x_w \mid \text{P})$
lunch	1	1/3	**1/3**	1/2	**1/2**
meeting	0	1/3	**2/3**	0	**1**
expenses	0	1/3	**2/3**	0	**1**
submitted	0	1/3	**2/3**	0	**1**
eom	1	1/3	**1/3**	1/2	**1/2**
meet	0	0	**1**	1/2	**1/2**
for	0	0	**1**	1/2	**1/2**
lol	0	1/3	**2/3**	1/2	**1/2**
UNKNOWN	0	0	**1**	0	**1**

$$
\begin{aligned}
p(M \mid \text{B}) &= \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot 1 \cdot 1 \cdot \frac{2}{3} \cdot 1 = \frac{16}{243} \approx .07 \\
p(M \mid \text{P}) &= \frac{1}{2} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot 1 = \frac{1}{32} = .03
\end{aligned}
$$

Naive Bayes Prediction for "Lunch? EOM"

- Input: "Lunch? EOM" $\Longrightarrow$ $M = \{\text{lunch}, \text{eom}\}$
- Message probability, conditional on message type:

$$
\begin{aligned}
p(M \mid \mathrm{B}) &= \frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3}\cdot 1\cdot 1\cdot\frac{2}{3}\cdot 1 = \frac{16}{243} \approx .07 \\
p(M \mid \mathrm{P}) &= \frac{1}{2}\cdot 1\cdot 1\cdot 1\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot 1 = \frac{1}{32} = .03
\end{aligned}
$$

- What does it mean that $p(M \mid \mathrm{P}) = .03$?
  - 3% of personal messages have same bag of words as M.

Naive Bayes Prediction

- Input: "Lunch? EOM" $\implies$ $M = \{\text{lunch}, \text{eom}\}$
- Output:

$$\begin{aligned} p(\text{BUSINESS} \mid M) &\propto p(\text{B})p(M \mid \text{B}) \\ &= \frac{3}{5} \cdot \frac{16}{243} = \frac{16}{405} \\ p(\text{PERSONAL} \mid M) &\propto p(\text{P})p(M \mid \text{P}) \\ &= \frac{2}{5} \cdot \frac{1}{32} = \frac{1}{90} \end{aligned}$$

- Now just renormalize:

$$\begin{aligned} p(\text{BUSINESS} \mid M) &= \frac{16}{405} \Big/ \left(\frac{1}{90} + \frac{16}{405} \right) \approx 0.78 \\ p(\text{PERSONAL} \mid M) &= \frac{1}{90} \Big/ \left(\frac{1}{90} + \frac{16}{405} \right) \approx 0.22 \end{aligned}$$

Naive Bayes Prediction: Issue With Zeros

- Input: $M =$"Meeting?"
- Output:

$$\begin{aligned} p(\text{BUSINESS} \mid M) &\propto \frac{1}{3} \\ p(\text{PERSONAL} \mid M) &\propto 0 \end{aligned}$$

- Renormalizing:

$$\begin{aligned} p(\text{BUSINESS} \mid M) &= 1 \\ p(\text{PERSONAL} \mid M) &= 0 \end{aligned}$$

- This is bad:
  - Never want to predict probability 0 if something is possible.
- Worse: Zero counts common for small sample sizes and rare features.

Laplace Smoothing

- **Laplace Smoothing** is a traditional fix to the 0 count issue.
- Idea is to add 1 to every empirical count:

$$\hat{p}(\text{lunch} \mid \text{PERSONAL}) = \frac{1 + \sum 1(\text{lunch and PERSONAL})}{1 + \sum 1(\text{PERSONAL})}$$

- The added 1 is called a **pseudocount**.
- Like assuming every outcome that can occur was observed at least once.
- Seems to solve the problem – but is there a more principled approach?

Bayesian Naive Bayes

- Be **Bayesian** and put a beta prior on each parameter.
- **Option 1:** Use posterior mean as point estimate for each parameter, then continue as before.
  - Laplace smoothing is a special case, in which priors are all $\text{Beta}(1,1)$.
- **Option 2: Go full Bayesian**.
  - No parameter estimates. Base everything on posterior $\theta \mid \mathcal{D}$.
- Predict with the predictive distribution:

$$y \mid x, \mathcal{D}$$

  - Recall, this is integrating out the parameter θ w.r.t. the posterior distribution.