Bayesian Classifiers, Conditional Independence and Naïve Bayes

Required reading:
“Naïve Bayes and Logistic Regression”
(available on class website)

Machine Learning 10-701
Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University
Jan 20, 2010

Announcements

• Homework 1 due today
• Homework 2 out soon – watch email

• Auditors must
  – officially register to audit
  – hand in at least n-1 or the n homeworks
Let’s learn classifiers by learning $P(Y|X)$

Suppose $Y=\text{Wealth}$, $X=\langle \text{Gender}, \text{HoursWorked} \rangle$

| Gender | HrsWorked | $P(\text{rich} | \text{G,HW})$ | $P(\text{poor} | \text{G,HW})$ |
|--------|-----------|-----------------|-----------------|
| F      | <40.5     | 0.09            | 0.91            |
| F      | >40.5     | 0.21            | 0.79            |
| M      | <40.5     | 0.23            | 0.77            |
| M      | >40.5     | 0.38            | 0.62            |

How many parameters must we estimate?

Suppose $X = \langle X_1, \ldots, X_n \rangle$

where $X_i$ and $Y$ are boolean RV’s

To estimate $P(Y|X_1, X_2, \ldots, X_n)$

If we have 30 $X_i$’s instead of 2?
Can we reduce params by using Bayes Rule?

Suppose $X = \langle X_1, \ldots, X_n \rangle$

where $X_i$ and $Y$ are boolean RV's

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i)P(Y = y_i)}{P(X = x_j)}$$

Equivalently:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i)P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k)P(Y = y_k)}$$
Naïve Bayes

Naïve Bayes assumes

\[ P(X_1 \ldots X_n | Y) = \prod_i P(X_i | Y) \]

i.e., that \( X_i \) and \( X_j \) are conditionally independent given \( Y \), for all \( i \neq j \)

Conditional Independence

Definition: \( X \) is conditionally independent of \( Y \) given \( Z \), if the probability distribution governing \( X \) is independent of the value of \( Y \), given the value of \( Z \)

\[ (\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k) \]

Which we often write

\[ P(X | Y, Z) = P(X | Z) \]

E.g.,

\[ P(\text{Thunder} | \text{Rain, Lightning}) = P(\text{Thunder} | \text{Lightning}) \]
Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given $Y$

Given this assumption, then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$

$$= P(X_1|Y)P(X_2|Y)$$

in general: $P(X_1...X_n|Y) = \prod_i P(X_i|Y)$

How many parameters to describe $P(X_1...X_n|Y)$? $P(Y)$?
• Without conditional indep assumption?
• With conditional indep assumption?

---

**Naïve Bayes in a Nutshell**

Bayes rule:

$$P(Y = y_k|X_1...X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

Assuming conditional independence among $X_i$’s:

$$P(Y = y_k|X_1...X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

So, classification rule for $X^{new} = <X_1, ..., X_n>$ is:

$$Y^{new} \leftarrow \text{arg max}_{y_k} P(Y = y_k)\prod_i P(X_i^{new}|Y = y_k)$$
Naïve Bayes Algorithm – discrete $X_i$

- Train Naïve Bayes (examples)
  
  for each* value $y_k$
  
  estimate $\pi_k \equiv P(Y = y_k)$
  
  for each* value $x_{ij}$ of each attribute $X_i$
  
  estimate $\theta_{ijk} \equiv P(X_i = x_{ij}|Y = y_k)$

- Classify ($X_{new}$)

  $Y_{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_{new}^i | Y = y_k)$

  $Y_{new} \leftarrow \arg \max_{y_k} \pi_k \prod_i \theta_{ijk}$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates (MLE’s):

$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$

$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}}$

Number of items in dataset D for which $Y=y_k$
Example: Live in Sq Hill? $P(S|G,D,M)$

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive to CMU
- $M=1$ iff Rachel Maddow fan

What probability parameters must we estimate?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S=1)$ :</td>
<td>$P(S=0)$ :</td>
</tr>
<tr>
<td>$P(D=1</td>
<td>S=1)$ :</td>
</tr>
<tr>
<td>$P(D=1</td>
<td>S=0)$ :</td>
</tr>
<tr>
<td>$P(G=1</td>
<td>S=1)$ :</td>
</tr>
<tr>
<td>$P(G=1</td>
<td>S=0)$ :</td>
</tr>
<tr>
<td>$P(M=1</td>
<td>S=1)$ :</td>
</tr>
<tr>
<td>$P(M=1</td>
<td>S=0)$ :</td>
</tr>
</tbody>
</table>
Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero. (e.g., $X_{373} = \text{Birthday\_Is\_January\_30\_1990}$)

- Why worry about just one parameter out of many?

- What can be done to avoid this?

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose $\theta$ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose $\theta$ that is most probable given prior probability and the data

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$
Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \mid Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Beta, Dirichlet priors):

$$\tilde{\pi}_k = \tilde{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + \alpha_k}{|D| + \sum_m \alpha_m}$$

$$\tilde{\theta}_{ijk} = \tilde{P}(X_i = x_{ij} \mid Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\} + \alpha'_k}{\#D\{Y = y_k\} + \sum_m \alpha'_m}$$

Only difference: “imaginary” examples

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose $\theta$ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg \max_\theta P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose $\theta$ that is most probable given prior probability and the data

$$\hat{\theta} = \arg \max_\theta P(\theta \mid \mathcal{D})$$

$$= \arg \max_\theta = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$
**Beta prior distribution – P(θ)**

\[ P(\theta) = \frac{\theta^{\beta_H-1}(1 - \theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T) \]

- **Likelihood function:** \( P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T} \)
- **Posterior:** \( P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \)

**MAP for Beta distribution**

\[ P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1 - \theta)^{\beta_T+\alpha_T-1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H+\alpha_H, \beta_T+\alpha_T) \]

- **MAP:** use most likely parameter:
  \[ \hat{\theta} = \arg \max_\theta P(\theta \mid \mathcal{D}) = \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!
Dirichlet distribution

- number of heads in N flips of a two-sided coin
  - follows a binomial distribution
  - Beta is a good prior (conjugate prior for binomial)

- what it's not two-sided, but k-sided?
  - follows a multinomial distribution
  - Dirichlet distribution is the conjugate prior

\[ P(\theta_1, \theta_2, ... \theta_K) = \frac{1}{\text{B}(\alpha)} \prod_{i=1}^{K} \theta_i^{(\alpha_i-1)} \]

Naïve Bayes: Subtlety #2

Often the \( X_i \) are not really conditionally independent

- We use Naïve Bayes in many cases anyway, and it often works pretty well
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- What is effect on estimated \( P(Y|X) \)?
  - Special case: what if we add two copies: \( X_i = X_k \)
Special case: what if we add two copies: $X_i = X_k$

Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?
Baseline: Bag of Words Approach

Learning to Classify Text

Target concept Interesting? : Document → {+, −}

1. Represent each document by vector of words
   • one attribute per word position in document

2. Learning: Use training examples to estimate
   • \( P(+) \)
   • \( P(−) \)
   • \( P(doc|+) \)
   • \( P(doc−) \)

Naive Bayes conditional independence assumption

\[
P(doc|v_j) = \prod_{i=1}^{\text{length}(doc)} P(a_i = w_k|v_j)
\]

where \( P(a_i = w_k|v_j) \) is probability that word in position \( i \) is \( w_k \), given \( v_j \)

one more assumption:

\( P(a_i = w_k|v_j) = P(a_{im} = w_k|v_j), \forall i, m \)
Twenty NewsGroups

Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

<table>
<thead>
<tr>
<th>comp.graphics</th>
<th>misc.forsale</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp.os.ms-windows.misc</td>
<td>rec.autos</td>
</tr>
<tr>
<td>comp.sys.ibm.pc.hardware</td>
<td>rec.motorcycles</td>
</tr>
<tr>
<td>comp.sys.mac.hardware</td>
<td>rec.sport.baseball</td>
</tr>
<tr>
<td>comp.windows.x</td>
<td>rec.sport.hockey</td>
</tr>
<tr>
<td>alt.atheism</td>
<td>sci.space</td>
</tr>
<tr>
<td>soc.religion.christian</td>
<td>sci.crypt</td>
</tr>
<tr>
<td>talk.religion.misc</td>
<td>sci.electronics</td>
</tr>
<tr>
<td>talk.politics.mideast</td>
<td>sci.med</td>
</tr>
<tr>
<td>talk.politics.misc</td>
<td></td>
</tr>
<tr>
<td>talk.politics.guns</td>
<td></td>
</tr>
</tbody>
</table>

Naive Bayes: 89% classification accuracy

Learning Curve for 20 Newsgroups

Accuracy vs. Training set size (1/3 withheld for test)
What if we have continuous $X_i$?

Eg., image classification: $X_i$ is $i^{th}$ pixel

Still have:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Just need to decide how to represent $P(X_i | Y)$
What if we have continuous $X_i$?

Eg., image classification: $X_i$ is $i^{th}$ pixel

Gaussian Naïve Bayes (GNB): assume

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma^2_{ik}}}$$

Sometimes assume variance
• is independent of $Y$ (i.e., $\sigma_i$),
• or independent of $X_i$ (i.e., $\sigma_k$)
• or both (i.e., $\sigma$)

Gaussian Naïve Bayes Algorithm – continuous $X_i$
(but still discrete $Y$)

• Train Naïve Bayes (examples)
  for each value $y_k$
    estimate* $\pi_k \equiv P(Y = y_k)$
    for each attribute $X_i$ estimate
      class conditional mean $\mu_{ik}$, variance $\sigma_{ik}$

• Classify ($X^{new}$)
  $$Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X^{new}_i \mid Y = y_k)$$
  $$Y^{new} \leftarrow \arg \max_{y_k} \pi_k \prod_i \text{Normal}(X^{new}_i, \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...
Estimating Parameters: $Y$ discrete, $X_i$ continuous

Maximum likelihood estimates:

\[
\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)
\]

\[
\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)
\]

**GNB Example:** Classify a person’s cognitive activity, based on brain image

- are they reading a sentence or viewing a picture?
- reading the word “Hammer” or “Apartment”
- viewing a vertical or horizontal line?
- answering the question, or getting confused?
Stimuli for our study:

- 60 distinct exemplars, presented 6 times each

fMRI voxel means for “bottle”: means defining $P(X_i \mid Y=\text{"bottle"})$

Mean fMRI activation over all stimuli:

“bottle” minus mean activation:
Rank Accuracy Distinguishing among 60 words

Tools vs Buildings: where does brain encode their word meanings?

Accuracies of cubical 27-voxel Naïve Bayes classifiers centered at each significant voxel [0.7-0.8]
What you should know:

• Training and using classifiers based on Bayes rule

• Conditional independence
  – What it is
  – Why it’s important

• Naïve Bayes
  – What it is
  – Why we use it so much
  – Training using MLE, MAP estimates
  – Discrete variables and continuous (Gaussian)

Questions:

• What is the error will classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?

• Can you use Naïve Bayes for a combination of discrete and real-valued $X_i$?

• How can we easily model just 2 of $n$ attributes as dependent?

• What does the decision surface of a Naïve Bayes classifier look like?