Naïve Bayes Classifier

Many of these slides are derived from Tom Mitchell, William Cohen, Eric Xing. Thanks!
Let’s learn classifiers by learning $P(Y|X)$

Consider $Y=$Wealth, $X=<$Gender, HoursWorked$>$

| gender | hours_worked | wealth | P(rich | G,HW) | P(poor | G,HW) |
|--------|--------------|--------|-----------|-----------|
| Female | v0:40.5-     | poor   | 0.253122  |           |
|        |              | rich   | 0.0245895 |           |
|        | v1:40.5+     | poor   | 0.0421768 |           |
|        |              | rich   | 0.0116293 |           |
| Male   | v0:40.5-     | poor   | 0.331313  |           |
|        |              | rich   | 0.0971295 |           |
|        | v1:40.5+     | poor   | 0.134106  |           |
|        |              | rich   | 0.105933  |           |

$P(\text{gender}, \text{hours}\_\text{worked}, \text{wealth})$ => $P(\text{wealth} | \text{gender}, \text{hours}\_\text{worked})$
How many parameters must we estimate?

Suppose $X = <X_1, \ldots, X_n>$
where $X_i$ and $Y$ are boolean RV’s

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

$2^n$ quantities need to be estimated!

If we have 30 boolean $X_i$’s: $P(Y | X_1, X_2, \ldots, X_{30})$

$2^{30} \sim 1$ billion!

You need lots of data or a very small $n$
Can we reduce params using Bayes Rule?

Suppose $X = \langle X_1, \ldots, X_n \rangle$
where $X_i$ and $Y$ are boolean RV’s

How many parameters for $P(X|Y) = P(X_1, \ldots, X_n | Y)$?

$(2^n-1) \times 2$

How many parameters for $P(Y)$?

1
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

• Two variables $A, B$ are independent if

\[ P(A \land B) = P(A) \times P(B) \]

\[ \forall a, b : P(A = a \land B = b) = P(A = a) \times P(B = b) \]

• Two variables $A, B$ are conditionally independent given $C$ if

\[ P(A, B \mid C) = P(A \mid C) \times P(B \mid C) \]

\[ \forall a, b, c : P(A = a \land B = b \mid C = c) = P(A = a \mid C = c) \times P(B = b \mid C = c) \]
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. \(P(Thunder | Rain, Lightning) = P(Thunder | Lightning)\)
Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given $Y$

Given this assumption, then:

\[
P(X_1, X_2 | Y) = P(X_1 | X_2, Y) P(X_2 | Y)
\]

\[
= P(X_1 | Y) P(X_2 | Y)
\]

\[
\text{Chain rule}
\]

\[
P(X_1 \ldots X_n | Y) = \prod_{i} P(X_i | Y)
\]

\[
(2^n-1) \times 2^{2n}
\]

\[
\text{Conditional Independence}
\]
Reducing the number of parameters to estimate

\[ P(Y|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y)P(Y)}{P(X_1,\ldots,X_n)} \]

To make this tractable we naively assume conditional independence of the features given the class: ie

\[ P(X_1,\ldots,X_n|Y) = P(X_1|Y) \cdot P(X_2|Y) \cdots P(X_n|Y) \]

Now: I only need to estimate \( \ldots \) parameters:

\[ P(X_1|Y), P(X_2|Y),\ldots,P(X_n|Y), P(Y) \]

How many parameters to describe \( P(X_1\ldots X_n|Y) \)? \( P(Y) \)?
- Without conditional indep assumption? \( (2^n-1) \times 2 + 1 \)
- With conditional indep assumption? \( 2n+1 \)
Naïve Bayes Algorithm – discrete $X_i$

- **Train Naïve Bayes** (given data for $X$ and $Y$) for each* value $y_k$
  
  estimate \[ \pi_k \equiv P(Y = y_k) \]

for each* value $x_{ij}$ of each attribute $X_i$

estimate \[ \theta_{ijk} \equiv P(X_i = x_{ij}|Y = y_k) \]
Training Naïve Bayes Classifier Using MLE

• From the data D, estimate class priors.
  – For each possible value of Y, estimate $Pr(Y=y_1)$, $Pr(Y=y_2)$, .... $Pr(Y=y_k)$
  – An MLE estimate:
    $$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

• From the data, estimate the conditional probabilities
  – If every $X_i$ has values $x_{i1}, ..., x_{ik}$
    • for each $y_i$ and each $X_i$ estimate $q(i,j,k) = Pr(X_i = x_{ij} | Y = y_i)$
    $$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in dataset D for which $Y=y_k$
Naïve Bayes Algorithm – discrete $X_i$

- **Train Naïve Bayes** (given data for $X$ and $Y$) 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** ($X^{new}$)
  
  $Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$

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

* probabilities must sum to 1, so need estimate only n-1 of these...
Example: Live in Sq Hill? $P(S|G,D,E)$

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive or Carpool to CMU
- $E=1$ iff Even # letters last name

What probability parameters must we estimate?

$P(S=1)$ : $P(S=0)$ :
$P(D=1 | S=1)$ : $P(D=0 | S=1)$ :
$P(D=1 | S=0)$ : $P(D=0 | S=0)$ :
$P(G=1 | S=1)$ : $P(G=0 | S=1)$ :
$P(G=1 | S=0)$ : $P(G=0 | S=0)$ :
$P(E=1 | S=1)$ : $P(E=0 | S=1)$ :
$P(E=1 | S=0)$ : $P(E=0 | S=0)$ :
Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero. (e.g., nobody in your sample has $X_i = <40.5$ and $Y=\text{rich}$)

- Why worry about just one parameter out of many?

$$P(X_1 \ldots X_n \mid Y) = \prod_{i} P(X_i \mid Y)$$

If one of these terms is 0...

- What can be done to avoid this?
Estimating Parameters

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

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

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

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

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

Maximum likelihood estimates:

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

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

MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \land Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

Only difference: “imaginary” examples
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: $X_i = X_k$

$$P(X_1 \ldots X_n|Y) = \prod_{i} P(X_i|Y)$$

Redundant terms
About Naïve Bayes

• Naïve Bayes is blazingly fast and quite robust!
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

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.
Learning to classify document: $P(Y|X)$
the “Bag of Words” model

- $Y$ discrete valued. e.g., Spam or not
- $X = <X_1, X_2, \ldots X_n> = \text{document}$

- $X_i$ is a random variable describing the word at position $i$ in the document
- possible values for $X_i$ : any word $w_k$ in English

- Document = bag of words: the vector of counts for all $w_k$’s
  – (like #heads, #tails, but we have more than 2 values)
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 $x_j$ of each attribute $X_i$
  estimate $\theta_{ijk} \equiv P(X_i = x_j | Y = y_k)$

- Classify ($X^{new}$)

  $Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X^{new}_i | Y = y_k)$

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

* Additional assumption: word probabilities are position independent
  $\theta_{ijk} = \theta_{mjk}$ for all $i, m$
MAP estimates for bag of words

MAP estimate for multinomial

\[
\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} (\beta_m - 1)}
\]

\[
\theta_{aardvark} = P(X_i = \text{aardvark}) = \frac{\# \text{ observed 'aardvark'} + \# \text{ hallucinated 'aardvark'} - 1}{\# \text{ observed words} + \# \text{ hallucinated words} - k}
\]

What $\beta$’s should we choose?
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
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