Naive Bayes

Nupur Chatterji, Kenny Marino, Colin White
Outline

- Bayes’ Rule
- Naive Bayes
- MLE/MAP
- Questions about HW1?
Bayes’ Rule

\[ P(A | B) = \frac{P(B | A)P(A)}{P(B)} \]

- \( P(A) \) is known as the “prior”
- \( P(A | B) \) is known as the “posterior”
Other Forms of Bayes’ Rule

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)} \]

\[ P(A \mid B \land X) = \frac{P(B \mid A \land X)P(A \land X)}{P(B \land X)} \]

- Law of Total Probability
Joint Distributions

<table>
<thead>
<tr>
<th>Temp (T)</th>
<th>Mood (M)</th>
<th>Go To Class (C)</th>
<th>Gets A</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Happy</td>
<td>Yes</td>
<td>.25</td>
</tr>
<tr>
<td>High</td>
<td>Happy</td>
<td>No</td>
<td>.05</td>
</tr>
<tr>
<td>High</td>
<td>Sad</td>
<td>Yes</td>
<td>.05</td>
</tr>
<tr>
<td>High</td>
<td>Sad</td>
<td>No</td>
<td>.15</td>
</tr>
<tr>
<td>Low</td>
<td>Happy</td>
<td>Yes</td>
<td>.20</td>
</tr>
<tr>
<td>Low</td>
<td>Happy</td>
<td>No</td>
<td>.05</td>
</tr>
<tr>
<td>Low</td>
<td>Sad</td>
<td>Yes</td>
<td>.15</td>
</tr>
<tr>
<td>Low</td>
<td>Sad</td>
<td>No</td>
<td>.10</td>
</tr>
</tbody>
</table>

We want to know $P(A|T, M, C)$

One solution:
Consider all combos of $2^3$ features separately
Joint Distributions

- Learning $P(Y|X)$ takes way too much data

Solution: Naive Bayes assumption

\[ P(X_1, X_2, ..., X_n|Y) = \prod P(X_i|Y) \]

$X_i, X_j$ are conditionally independent given $Y$
Conditional Independence

- Height is not independent of weight
- Height is independent of weight, given age

X is conditionally independent of Y given Z if

For all \(x,y,z,\) \(P(X=x|Y=y,Z=z) = P(X=x|Z=z)\)
Naive Bayes Algorithm

Given training data, features $X_1, \ldots, X_n$, label $Y$

Given a new datapoint $X_{\text{new}} = X_{1_{\text{new}}}, \ldots, X_{n_{\text{new}}}$

Want to compute most probable label

$Y_{\text{new}} = \arg\max_Y P(Y = y | X_{1_{\text{new}}}, \ldots, X_{n_{\text{new}}})$

$Y_{\text{new}} = \arg\max_{y_k} P(Y = y_k) \prod_i P(X_{i_{\text{new}}} | Y = y_k)$
Naive Bayes Algorithm (Binary features & label)

Given training data, feats $X_1, \ldots, X_n$, label $Y$

- Estimate $\pi_1 = P(Y=1)$
- For each $x_i$, estimate $\theta_{i,1} = P(X_i = 1 \mid Y=1)$

Given test set,

- Predict 1 if $\pi_1 \prod_i \theta_i > 0.5$
Maximum Likelihood Estimate

Estimate $\pi_i$

- $P(Y=1) = \frac{\# Y=1}{n}$

Estimate $\theta_i$

- $P(X_i = 1 \mid Y = 1) = \frac{\# X_i=1 \text{ and } Y = 1}{\# Y = 1}$
Maximum A Priori

- Recall the prediction rule $\pi_i \Pi_i \theta_i > 0.5$
- If some $\theta_i = P(X_i = 1 \mid Y=1)$ is zero, then we would predict $Y=0$ for all datapoints s.t. $X_i = 1$ (why?)

Solution: use MAP (hallucinated examples)

- $\theta_i = \frac{[\text{# } X_i=1 \text{ and } Y = 1]+c}{[\text{# } Y = 1]+2c}$
Naive Bayes Assumption

- Very likely, the Naive Bayes assumption is not true

\[ P(X_1, X_2, ..., X_n | Y) = \prod_i P(X_i | Y) \]

- Still, Naive Bayes does well in real life
- Examples where it is extremely not satisfied vs. almost satisfied?
Any other questions?