Outline

• Conditional independence

• Naïve Bayes assumption and its consequences
  – Which (and how many) parameters must be estimated under different generative models (different forms for $P(X|Y)$ )
  • and why this matters

• How to train Naïve Bayes classifiers
  – MLE and MAP estimates
  – with discrete and/or continuous inputs $X_i$
Conditional Independence

• Given random variables X, Y and Z.
• X is conditionally independent of Y given Z, if and only if:

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

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

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

If we have n variables, assuming conditional independence, we can write:

\[ P(X_1, X_2, ..., X_n | Y) = \prod_{i=1}^{n} P(X_i | Y) \]
# Parameters needed

- How many parameters we need to estimate:

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

where \( X_i \) and \( Y \) are Boolean random variables

Without conditional independence assumption?

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

With conditional independence assumption?

\[ 2n \]
Suppose we have $n$ variables

Conditionally Independent

$$P(X_1, X_2, \ldots, X_n|Y) = \prod_{i} P(X_i|Y)$$
Naïve Bayes in a Nutshell

Bayes rule:

\[ P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k)P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1 \ldots X_n | Y = y_j)} \]

Assuming conditional independence among \( X_i \)'s:

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

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

\[ Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k) \]
Naïve Bayes Algorithm

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 $X^{new}$

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

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$
• In class we constructed a Naïve Bayes classifier to predict if a someone lives in Squirrel Hill based on variables such as:
  – Driving to CMU
  – Shop at Giant Eagle
  – Even # letters in last name
• Note that we used all discrete variables
• We are not limited to discrete $X_i$!
What if we have continuous $X_i$?

Gaussian Naïve Bayes (GNB): assume

$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2}(\frac{x - \mu_{ik}}{\sigma_{ik}})^2}$$

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 Distribution
(also called “Normal”)

$p(x)$ is a probability density function, whose integral (not sum) is 1.

\[ p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \]

The probability that $X$ will fall into the interval $(a, b)$ is given by
\[ \int_a^b p(x) \, dx \]

- Expected, or mean value of $X$, $E[X]$, is
  \[ E[X] = \mu \]

- Variance of $X$ is
  \[ Var(X) = \sigma^2 \]

- Standard deviation of $X$, $\sigma_X$, is
  \[ \sigma_X = \sigma \]
Gaussian Naïve Bayes Algorithm

for each value \( y_k \):

\[ \pi_k \equiv P(Y = y_k) \]

For each attribute \( X_i \):

\[ P(X_i | Y = y_k) \]

class conditional mean and variance \( \mu_{ik}, \sigma_{ik} \)

Classify \( X^{\text{new}} \)

\[ Y^{\text{new}} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{\text{new}} | Y = y_k) \]

\[ Y^{\text{new}} \leftarrow \arg\max_{y_k} \pi_k \prod_i \mathcal{N}(X_i^{\text{new}}; \mu_{ik}, \sigma_{ik}) \]
<table>
<thead>
<tr>
<th>IsPopular</th>
<th>Daily Tweets</th>
<th>Has Facebook?</th>
<th>Years left to graduation</th>
<th>ML Grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td><strong>85</strong></td>
<td>0</td>
<td>2</td>
<td><strong>85</strong></td>
</tr>
<tr>
<td>no</td>
<td>80</td>
<td>1</td>
<td>2</td>
<td>90</td>
</tr>
<tr>
<td>yes</td>
<td><strong>83</strong></td>
<td>0</td>
<td>3</td>
<td><strong>86</strong></td>
</tr>
<tr>
<td>yes</td>
<td>70</td>
<td>0</td>
<td>1</td>
<td>96</td>
</tr>
<tr>
<td>yes</td>
<td><strong>68</strong></td>
<td>0</td>
<td>1</td>
<td><strong>80</strong></td>
</tr>
<tr>
<td>no</td>
<td>65</td>
<td>1</td>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td>yes</td>
<td><strong>64</strong></td>
<td>1</td>
<td>3</td>
<td><strong>65</strong></td>
</tr>
<tr>
<td>no</td>
<td>72</td>
<td>0</td>
<td>2</td>
<td>95</td>
</tr>
<tr>
<td>yes</td>
<td><strong>69</strong></td>
<td>0</td>
<td>2</td>
<td><strong>70</strong></td>
</tr>
<tr>
<td>yes</td>
<td>75</td>
<td>0</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>yes</td>
<td><strong>75</strong></td>
<td>1</td>
<td>2</td>
<td><strong>70</strong></td>
</tr>
<tr>
<td>yes</td>
<td>72</td>
<td>1</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>yes</td>
<td><strong>81</strong></td>
<td>0</td>
<td>3</td>
<td><strong>75</strong></td>
</tr>
<tr>
<td>no</td>
<td>71</td>
<td>1</td>
<td>1</td>
<td>91</td>
</tr>
</tbody>
</table>
Naïve Bayes Classification Rule

\[
Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_{i} P(X_i^{new} | Y = y_k)
\]

New example

Classify:

\[
X^{new} = < \text{lefttoGraduation} = 3, \text{DailyTweets} = 60, \text{HasFacebook} = 0, \text{MLGrade} = 62>
\]
For continuous variables assume Gaussian Distribution

- \( P(\text{Daily Tweets} \mid \text{IsPopular}) = N(\mu, \sigma^2) \)

\[
P(X_i = x_{ij} \mid Y = y_k) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} \exp\left(\frac{-\left(x_{ij} - \mu_{ik}\right)^2}{2\sigma_{ik}^2}\right)
\]

\(i^{th}\) variable taking the \(x_{ij}\)  
Class variable \(Y\) taking the \(y_k\)

Variance depends on class variable and \(X_i\)

Mean depends on class variable and \(X_i\)
Training

Statistics 101

\[
\hat{\mu} = \frac{\sum_{l=1}^{D} X_l}{D}
\]

\[
\hat{\sigma}^2 = \frac{\sum_{l=1}^{D} (X_l - \mu)}{D - 1}
\]

\[\mu\_dailytweets\_yes=? \quad \sigma\_dailytweets\_yes=?\]

\[\mu\_dailytweets\_no=? \quad \sigma\_dailytweets\_no=?\]

\[\mu\_MLgrades\_yes=? \quad \sigma\_MLgrades\_yes=?\]

\[\mu\_MLgrades\_no=? \quad \sigma\_MLgrades\_no=?\]
Training

Statistics 101

\[ \hat{\mu} = \frac{\sum_{l=1}^{D} X_l}{D} \]

\[ \hat{\sigma}^2 = \frac{\sum_{l=1}^{D} (X_l - \mu)}{D - 1} \]

\[
\begin{align*}
\mu_{\text{dailytweets\_yes}} &= 73 & \sigma_{\text{dailytweets\_yes}} &= 6.2 \\
\mu_{\text{dailytweets\_no}} &= 74.6 & \sigma_{\text{dailytweets\_no}} &= 8 \\
\mu_{\text{MLgrades\_yes}} &= 79.1 & \sigma_{\text{MLgrades\_yes}} &= 10.2 \\
\mu_{\text{MLgrades\_no}} &= 86.2 & \sigma_{\text{MLgrades\_no}} &= 9.7
\end{align*}
\]
$$X_{new} = < lefttograduation = 3, \text{DailyTweets} = 60, \text{HasFacebook} = 0, \text{MLGrade} = 62 >$$

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\text{IsPopular} = \text{yes})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P(\text{HasFacebook} = 0</td>
<td>\text{IsPopular} = \text{yes})$</td>
<td></td>
</tr>
<tr>
<td>$P(\text{HasFacebook} = 1</td>
<td>\text{IsPopular} = \text{no})$</td>
<td></td>
</tr>
<tr>
<td>$P(\text{YearsLeft} = 1</td>
<td>\text{IsPopular} = \text{yes})$</td>
<td></td>
</tr>
<tr>
<td>$P(\text{YearsLeft} = 2</td>
<td>\text{IsPopular} = \text{yes})$</td>
<td></td>
</tr>
<tr>
<td>$P(\text{YearsLeft} = 3</td>
<td>\text{IsPopular} = \text{yes})$</td>
<td></td>
</tr>
<tr>
<td>$P(\text{DailyTweets} = 60</td>
<td>\text{IsPopular} = \text{yes})$</td>
<td></td>
</tr>
<tr>
<td>$P(\text{MLGrade} = 62</td>
<td>\text{IsPopular} = \text{yes})$</td>
<td></td>
</tr>
</tbody>
</table>
\[ X^{\text{new}} = \langle \text{lefttoGraduation} = 3, \text{DailyTweets} = 60, \text{HasFacebook} = 0, \text{MLGrade} = 62 \rangle \]

\[
\begin{align*}
P(\text{IsPopular} = \text{yes}) &= 0.643 \\
P(\text{IsPopular} = \text{no}) &= 0.357 \\
P(\text{HasFacebook} = 0 | \text{IsPopular} = \text{yes}) &= 0.667 \\
P(\text{YearsLeft} = 3 | \text{IsPopular} = \text{yes}) &= 0.286 \\
P(\text{DailyTweets} = 60 | \text{IsPopular} = \text{yes}) &= 0.0071 \\
P(\text{MLGrade} = 62 | \text{IsPopular} = \text{yes}) &= 0.0096 \\
P(\text{HasFacebook} = 0 | \text{IsPopular} = \text{yes}) &= 0.667 \\
P(\text{YearsLeft} = 3 | \text{IsPopular} = \text{no}) &= 0 \\
P(\text{DailyTweets} = 60 | \text{IsPopular} = \text{no}) &= 0.0094 \\
P(\text{MLGrade} = 60 | \text{IsPopular} = \text{no}) &= 0.0018
\end{align*}
\]
A lot of problems...

- In the training data, there are no yearsLeftGraduation=3|IsPopular=No
- What if the test example has yearsLeftGraduation=4
- It is not in the training data!
- We need to **smooth** or **regularize** the estimates to avoid overfitting training dataset
• MLE estimates

$$\theta_{MLE} = \arg\max_{\theta} P(X|\theta)$$

• MAP estimates

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|X)$$

$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta)$$

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

Prior to avoid 0 probabilities
Text Classification

• Y discrete valued.
  – e.g., Spam or not
• $X = \langle X_1, X_2, \ldots, X_n \rangle = \text{document}$

• $X_i$ is a random variable describing…

Answer 2:
• $X_i$ represents the $i^{th}$ word position in document
• $X_1 = \text{"I"}$, $X_2 = \text{"am"}$, $X_3 = \text{"pleased"}$
• and, let’s assume the $X_i$ are iid (indep, identically distributed)
The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nation's pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC. Delegates to the three-day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said. A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter
Questions
Acknowledgements

- Some slides are taken from Tom Mitchell's 10-601 lecture notes