Today:
- Naïve Bayes
  - discrete-valued $X_i$’s
  - Document classification
- Gaussian Naïve Bayes
  - real-valued $X_i$’s
  - Brain image classification
- Form of decision surfaces

Readings:

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

Optional
- Bishop 1.2.4
- Bishop 4.2

Naïve Bayes in a Nutshell

Bayes rule:

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

Assuming conditional independence among $X_i$’s: \(i \neq j\)

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

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

\[Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)\]
Another way to view Naïve Bayes (Boolean Y):

Decision rule: is this quantity greater or less than 1?

\[
1 \geq \frac{P(Y = 1|X_1 \ldots X_n)}{P(Y = 0|X_1 \ldots X_n)} = \frac{P(Y = 1) \prod_i P(X_i|Y = 1)}{P(Y = 0) \prod_i P(X_i|Y = 0)}
\]

\[
0 \leq \log \frac{P(Y = 1|X_1 \ldots X_n)}{P(Y = 0|X_1 \ldots X_n)} = \log \frac{P(Y = 1)}{P(Y = 0)} + \sum_i \log \frac{P(X_i|Y = 1)}{P(X_i|Y = 0)}
\]

\[
\Theta_{ik} = \frac{P(X_i = 1|Y = k)}{P(X_i = 0|Y = k)} \quad D \geq \log \frac{P(Y = 1)}{P(Y = 0)} + \sum_i \log \frac{\Theta_{ik}}{\Theta_{i0}} + \sum_i (I - \Theta_{ik}) \log \frac{(I - \Theta_{ik})}{(I - \Theta_{i0})}
\]

\[
P(S | D,G,M)
\]
Naïve Bayes: classifying text documents

• Classify which emails are spam?
• Classify which emails promise an attachment?

How shall we represent text documents for Naïve Bayes?

Learning to classify documents: $P(Y|X)$

• $Y$ discrete valued.
  – e.g., Spam or not
• $X = <X_1, X_2, \ldots X_n>$ = document

• $X_i$ is a random variable describing…
Learning to classify documents: $P(Y|X)$

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

- $X_i$ is a random variable describing...
  Answer 1: $X_i$ is boolean, 1 if word $i$ is in document, else 0
  e.g., $X_{\text{pleased}} = 1$

Issues?

Learning to classify documents: $P(Y|X)$

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

- $X_i$ is a random variable describing...
  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)
  $$P(X_i|Y) = P(X_j|Y) \quad (\forall i, j)$$
Learning to classify document: $P(Y|X)$
the “Bag of Words” model

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

• $X_i$ are iid random variables. Each represents the word at its position $i$ in the document
• Generating a document according to this distribution = rolling a 50,000 sided die, once for each word position in the document

• The observed counts for each word follow a ??? distribution

---

**Multinomial Distribution**

• $P(\theta)$ and $P(\theta|D)$ have the same form

---

**Eg. 2** Dice roll problem ($k$ outcomes instead of 2)

Likelihood is $\sim$ Multinomial $\{\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}\}$

$$P(D|\theta) = \theta_1^{\text{count for side 1}} \cdot \theta_2^{\text{count for side 2}} \ldots \cdot \theta_k^{\text{count for side k}}$$

If prior is Dirichlet distribution,

$$P(\theta) = \pi \frac{\prod_{i=1}^{k} \theta_i^{\beta_i-1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

**For Multinomial, conjugate prior is Dirichlet distribution.**
Multinomial Bag of Words

MAP estimates for bag of words

Map estimate for multinomial

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

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

What $\beta$’s should we choose?
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 ($X_{new}^*$)
  
  $Y_{new}^* \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_{i,new}^*|Y = y_k)$

  prob that word $x_{ij}$ appears in position $i$, given $Y=y_k$

* Additional assumption: word probabilities are position independent
  $\theta_{ijk} = \theta_{mjk}$ for $i \neq m$

---

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
For code and data, see www.cs.cmu.edu/~tom/mlbook.html click on "Software and Data"

What if we have continuous $X_i$?

Eg., image classification: $X_i$ is real-valued $i^{th}$ pixel
What if we have continuous $X_i$?

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

Naïve Bayes requires $P(X_i \mid Y=y_k)$, but $X_i$ is real (continuous)

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

Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution

Gaussian Distribution
(also called “Normal”)

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

$$p(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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

$$\sigma_X = \sigma$$
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., $\sigma_k$)

Train Naïve Bayes (examples)
- for each value $y_k$
- estimate $P(Y = y_k)$
- for each attribute $X_i$
- estimate $P(X_i | Y = y_k)$
  - class conditional mean $\mu_{ik}$
  - variance $\sigma_{ik}$

Classify ($X_{new}$)

$$Y_{new} \leftarrow \text{arg max}_{y_k} P(Y = y_k) \prod_i P(X_{i new} | Y = y_k)$$

$$Y_{new} \leftarrow \text{arg max}_{y_k} \pi_k \prod_i \mathcal{N}(X_{i new}, \mu_{ik}, \sigma_{ik})$$

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

Maximum likelihood estimates:

$$
\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X^j_i \delta(Y^j = y_k)
$$

$$
\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X^j_i - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)
$$

How many parameters must we estimate for Gaussian Naïve Bayes if $Y$ has $k$ possible values, $X=<X_1, \ldots, X_n>$?

$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{ik}} e^{-\frac{(x_i - \hat{\mu}_{ik})^2}{2\hat{\sigma}_{ik}^2}}$
What is form of decision surface for Gaussian Naïve Bayes classifier?

eg., if we assume attributes have same variance, indep of $Y$ 
($\sigma_{ik} = \sigma$)

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

• reading a sentence or viewing a picture?
• reading the word describing a “Tool” or “Building”?
• answering the question, or getting confused?
Mean activations over all training examples for Y="bottle"

Y is the mental state (reading “house” or “bottle”)
Xᵢ are the voxel activities,
this is a plot of the μ's defining P(Xᵢ | Y="bottle")

\[ P(Y|X) \propto P(Y) \prod P(X_i|Y) \]

Classification task: is person viewing a “tool” or “building”?

Classification accuracy

Statistically significant
p<0.05
Where is information encoded in the brain?

Accuracies of cubical 27-voxel classifiers centered at each significant voxel [0.7-0.8]

Naïve Bayes: What you should know

• Designing classifiers based on Bayes rule

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

• 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 $X_i$
Questions to think about:

• 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?

• How would you select a subset of $X_i$'s?