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

Recently:
- Bayes classifiers to learn $P(Y|X)$
- MLE and MAP estimates for parameters of $P$
- Conditional independence
- Naïve Bayes → make Bayesian learning practical

Next:
- Text classification
- Naïve Bayes and continuous variables $X_i$:
  - Gaussian Naïve Bayes classifier
- Learn $P(Y|X)$ directly
  - Logistic regression, Regularization, Gradient ascent
- Naïve Bayes or Logistic Regression?
  - Generative vs. Discriminative classifiers
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 \(x^{\text{new}} = < x_1, \ldots, x_n >\) is:

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

Another way to view Naïve Bayes (Boolean Y):

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

\[
\frac{P(Y = 1 | X_1 \ldots X_n)}{P(Y = 0 | X_1 \ldots X_n)} = \frac{P(Y = 1) \prod_i P(X_i | Y = 1)}{P(Y = 0) \prod_i P(X_i | Y = 0)}
\]
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, ... 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 \( i^{\text{th}} \) 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 (6 outcomes instead of 2)

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

$$P(D|\theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

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

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

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>0</td>
</tr>
<tr>
<td>about</td>
<td>2</td>
</tr>
<tr>
<td>all</td>
<td>2</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
</tr>
<tr>
<td>apple</td>
<td>0</td>
</tr>
<tr>
<td>anxious</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>gas</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>oil</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Zaire</td>
<td>0</td>
</tr>
</tbody>
</table>

MAP estimates for bag of words

Map estimate for multinomial

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

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

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 $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$
  prob that word $x_{ij}$ appears in position $i$, given $Y=y_k$

- Classify ($X^{new}$)
  \[
  Y^{new} \leftarrow \arg \max_{y_k} \, P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k) \]
  \[
  Y^{new} \leftarrow \arg \max_{y_k} \, \pi_k \prod_i \theta_{ijk} \]

* 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

```plaintext
comp.graphics  comp.os.ms-windows.misc  misc.forsale
comp.sys.ibm.pc.hardware  comp.sys.mac.hardware  rec.autos
comp.windows.x  alt.atheism  rec.cars
soc.religion.christian  talk.religion.misc  rec.motorcycles
talk.politics.mideast  talk.politics.misc  rec.sport.baseball
talk.politics.guns  sci.space  rec.sport.hockey
sci.crypt  sci.electronics  sci.med
```
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)

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

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., $\sigma$),
- 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 $P(X_i|Y = y_k)$
  - class conditional mean $\mu_{ik}$, variance $\sigma_{ik}$

- Classify ($X_{\text{new}}$)
  $$Y_{\text{new}} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{\text{new}}|Y = y_k)$$
  $$Y_{\text{new}} \leftarrow \arg \max_{y_k} \pi_k \prod_i \mathcal{N}(X_i^{\text{new}}, \mu_{ik}, \sigma_{ik})$$

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

Maximum likelihood estimates:

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

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

How many parameters must we estimate for Gaussian Naïve Bayes if $Y$ has $k$ possible values, $X=$<X1, ..., Xn>?

$$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{ik}^2}} e^{-\frac{(x - \hat{\mu}_{ik})^2}{2\hat{\sigma}_{ik}^2}}$$
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?

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

Mean activations over all training examples for Y=“bottle”
Classification task: is person viewing a “tool” or “building”?

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?
Logistic Regression

Idea:
• Naïve Bayes allows computing $P(Y|X)$ by learning $P(Y)$ and $P(X|Y)$

• Why not learn $P(Y|X)$ directly?
• Consider learning \( f: X \rightarrow Y \), where
  
  • \( X \) is a vector of real-valued features, \(< X_1, ..., X_n >\)
  
  • \( Y \) is boolean
  
  • assume all \( X_i \) are conditionally independent given \( Y \)
  
  • model \( P(X_i | Y = y_k) \) as Gaussian \( N(\mu_{ik}, \sigma_i) \)
  
  • model \( P(Y) \) as Bernoulli \( (\pi) \)

• What does that imply about the form of \( P(Y|X) \)?

\[
P(Y = 1|X = < X_1, ..., X_n>) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}
\]

---

Derive form for \( P(Y|X) \) for continuous \( X_i \)

\[
P(Y = 1|X) = \frac{P(Y = 1)P(X|Y = 1)}{P(Y = 1)P(X|Y = 1) + P(Y = 0)P(X|Y = 0)}
\]

\[
= \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}
\]

\[
= \frac{1}{1 + \exp(\ln \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}
\]

\[
P(z | y_k) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(z - \mu_{ik})^2}{2\sigma_i^2}}
\]

\[
P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}
\]
Very convenient!

\[ P(Y = 1|X = < X_1, ...X_n >) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)} \]
implies
\[ P(Y = 0|X = < X_1, ...X_n >) = \]
implies
\[ \frac{P(Y = 0|X)}{P(Y = 1|X)} = \]
implies
\[ \ln \frac{P(Y = 0|X)}{P(Y = 1|X)} = \]

Very convenient!

\[ P(Y = 1|X = < X_1, ...X_n >) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)} \]
implies
\[ P(Y = 0|X = < X_1, ...X_n >) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)} \]
implies
\[ \frac{P(Y = 0|X)}{P(Y = 1|X)} = \exp(w_0 + \sum_i w_i X_i) \]
implies
\[ \ln \frac{P(Y = 0|X)}{P(Y = 1|X)} = w_0 + \sum_i w_i X_i \]