Classification

- Representing data:

- Choosing hypothesis class

- Learning: $h: X \mapsto Y$
  - $X$ – features
  - $Y$ – target classes
Suppose you know the following...

- Class-specific Dist.: \( P(X|Y) \)
  
  \[
  p(X | Y = 1) = p_1(X; \mu_1, \Sigma_1) \\
  p(X | Y = 2) = p_2(X; \mu_2, \Sigma_2)
  \]

- Bayes classifier:

- Class prior (i.e., "weight"): \( P(Y) \)

- This is a generative model of the data!

Recall Bayes Rule

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

Which is shorthand for:

\[
(\forall i, j) P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{P'(X = x_j)}
\]

Equivalently:

\[
(\forall i, j) P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{\sum_k P(X = x_j|Y = y_k)P(Y = y_k)}
\]
Recall Bayes Rule

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

Which is shorthand for:

\[(\forall i, j)P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i)P(Y = y_i)}{P(X = x_j)}\]

Common abbreviation:

\[(\forall i, j)P(y_i | x_j) = \frac{P(x_j | y_i)P(y_i)}{\sum_k P(x_j | y_k)P(y_k)}\]

Optimal classification

- **Theorem**: Bayes classifier is optimal!
  - That is
  \[ \text{error}_{true}(h_{Bayes}) \leq \text{error}_{true}(h), \forall h(x) \]

- **Proof**: 
  - Recall last lecture:

- How to learn a Bayes classifier?
Learning Bayes Classifier

- Training data:

  ![Training data diagram]

  - Learning = estimating $P(X|Y)$, and $P(Y)$
  - Classification = using Bayes rule to calculate $P(Y | X_{new})$

How hard is it to learn the optimal classifier?

- How do we represent these? How many parameters?
  - Prior, $P(Y)$:
    - Suppose $Y$ is composed of $k$ classes
  - Likelihood, $P(X|Y)$:
    - Suppose $X$ is composed of $n$ binary features

- Complex model $\rightarrow$ High variance with limited data!!!
Naïve Bayes:

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

assuming that \( X_i \) and \( X_j \) are conditionally independent given \( Y \), for all \( i \neq j \)

---

Conditional Independence

- \( X \) is conditionally independent of \( Y \) given \( Z \), if the probability distribution governing \( X \) is independent of the value of \( Y \), given the value of \( Z \)

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

Which we often write

\[ P(X \mid Y, Z) = P(X \mid Z) \]

- e.g.,

\[ P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning}) \]

- Equivalent to:

\[ P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z) \]
The Naïve Bayes assumption

- Naïve Bayes assumption:
  - Features are conditionally independent given class:
    \[ P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y) = P(X_1|Y)P(X_2|Y) \]
  - More generally:
    \[ P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y) \]

- How many parameters now?
  - Suppose \( X \) is composed of \( n \) binary features

The Naïve Bayes Classifier

- Given:
  - Prior \( P(Y) \)
  - \( n \) conditionally independent features \( X \) given the class \( Y \)
    - For each \( X_i \), we have likelihood \( P(X_i|Y) \)

- Decision rule:
  \[
  y^* = h_{NB}(x) = \arg \max_y P(y)P(x_1, \ldots, x_n | y) = \arg \max_y P(y) \prod_i P(x_i|y)
  \]

- If assumption holds, NB is optimal classifier!
Naïve Bayes Algorithm

- 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 ($X_{new}$)

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

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

Case Study: Text classification

- Classify e-mails
  - $Y = \{\text{Spam}, \text{NotSpam}\}$

- Classify news articles
  - $Y = \{\text{what is the topic of the article?}\}$

- Classify webpages
  - $Y = \{\text{Student, professor, project, …}\}$

- What about the features $X$?
  - The text!
Features $X$ are entire document – $X_i$ for $i^{th}$ word in article

Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opini
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
mediocre goaltender. A better choice would be
Tomas Sandstrom, though not through any fault of
his own, but because some thugs in Toronto decided

NB for Text classification

- $P(X|Y)$ is huge!!!
  - Article at least 1000 words, $X=\{X_1, \ldots, X_{1000}\}$
  - $X_i$ represents $i^{th}$ word in document, i.e., the domain of $X_i$ is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.

- NB assumption helps a lot!!!
  - $P(X=x_i|Y=y)$ is just the probability of observing word $x_i$ in a document on topic $y$

$$h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i|y)$$
Bag of words model

- Typical additional assumption – **Position in document doesn’t matter**: \( P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y) \)
  - “Bag of words” model – order of words on the page ignored
  - Sounds really silly, but often works very well!

\[
P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i|y) \quad \text{or} \quad P(y) \prod_{k=1}^{\text{LengthVol}} P(w_k|y)
\]

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

Bag of words model

- Typical additional assumption – **Position in document doesn’t matter**: \( P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y) \)
  - “Bag of words” model – order of words on the page ignored
  - Sounds really silly, but often works very well!

\[
P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i|y) \quad \text{or} \quad P(y) \prod_{k=1}^{\text{LengthVol}} P(w_k|y)
\]

In is lecture lecture next over person remember room sitting the the the to to up wake when you
Bag of Words Approach

- Learning phase:
  - Prior $P(Y)$
    - Count how many documents you have from each topic (+ prior)
  - $P(X|Y)$
    - For each topic, count how many times you saw word in documents of this topic (+ prior)

- Test phase:
  - For each document $x_{\text{new}}$
    - Use naïve Bayes decision rule

$$h_{NB}(x_{\text{new}}) = \arg \max_y P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i^{\text{new}}|y)$$

NB with Bag of Words for text classification

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

We draw on our proven strength from our first-proven oil and gas reserves. Our strategy emphasizes our natural gas business in a rapidly expanding market.

Our expanding refining and marketing operations in Africa and the Mediterranean are complemented already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

- aardvark 0
- about 2
- all 2
- Africa 1
- apple 1
- anxious 0
- gas 1
- oil 1
- Zaire 0
Learning NB: parameter estimation

- 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

\[
\hat{\theta} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\max_{\theta} \frac{P(\mathcal{D}|\theta)p(\theta)}{P(\mathcal{D})}
\]

- Bayesian estimate:

\[
\hat{\theta} = \int \theta p(\theta|\mathcal{D})d\theta
\]

MLE for the parameters of NB

Discrete features:

- Maximum likelihood estimates (MLE’s): \( \hat{\theta} = \arg\max_{\theta} P(\mathcal{D}|\theta) \)
- Given dataset
  - \( \text{Count}(A=a,B=b) \leftarrow \text{number of examples where } A=a \text{ and } B=b \)
Experimental Setup

- Dataset
  - 20 News Groups (20 classes)
  - 61,118 words, 18,774 documents
  - Class labels descriptions

<table>
<thead>
<tr>
<th>comp.graphics</th>
<th>rec.autos</th>
<th>sci.crypt</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp.os.ms-windows</td>
<td>rec.motorcycles</td>
<td>sci.electronics</td>
</tr>
<tr>
<td>comp.sys.ibm.pc.hardware</td>
<td>rec.sport.baseball</td>
<td>sci.med</td>
</tr>
<tr>
<td>comp.sys.mac.hardware</td>
<td>rec.sport.hockey</td>
<td>sci.space</td>
</tr>
<tr>
<td>misc.forsale</td>
<td>talk.politics.misc</td>
<td>talk.religion.misc</td>
</tr>
<tr>
<td></td>
<td>talk.politics.guns</td>
<td>alt.atheism</td>
</tr>
<tr>
<td></td>
<td>talk.politics.mideast</td>
<td>soc.religion.christian</td>
</tr>
</tbody>
</table>

Experimental Setup (con’d)

- Parameters
  - Binary feature;
  - Using MLE to estimate parameters

- Training/Test Sets:
  - Change training set (from 5% to 95%), treat remaining as test set
  - 10 runs
  - report average results

- Evaluation Criteria:
  \[
  \text{Accuracy} = \frac{\sum_{\text{test set}} I(\text{predict}, \text{true label})}{\# \text{ of test samples}}
  \]
Results: Binary Classes

- *alt.atheism* vs. *comp.graphics*
- *comp.windows.x* vs. *rec.motorcycles*
- *rec.autos* vs. *rec.sport.baseball*

### Results: Multiple Classes

- 5-out-of-20 classes
- 10-out-of-20 classes
- All 20 classes
Subtleties of NB classifier 1 –
Violating the NB assumption

- 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])
  - But the resulting probabilities \( P(Y|X_{\text{new}}) \) are biased toward 1 or 0 (why?)

Subtleties of NB classifier 2 –
Insufficient training data

- What if you never see a training instance where \( w_{1000} > 0 \) when \( Y=b \)?
  - e.g., \( Y=\{\text{SpamEmail or not} \} \), \( w_{1,999} = \{\text{‘pill’, ‘enhancement’, ‘Rolex’, ‘enlarge’ … } \} \)
  - \( P(\text{enlargement}>0 | Y=T) = 0 \)

- Thus, no matter what the values \( w_1, \ldots, w_n/\text{enlargement’ take:} \)
  - \( P(Y=T | w_1, w_2, \ldots, \text{enlargement, …, } w_n) = 0 \)

- What now???
MAP for the parameters of NB

Discrete features:

- Maximum a Posteriori (MAP) estimate: (MAP’s): 
  \[
  \hat{\theta} = \arg \max_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)}
  \]

- Given prior:
  - Consider binary feature
  - \( \theta \) is a Bernoulli rate

Bayesian learning for NB parameters – a.k.a. smoothing

- Posterior distribution of \( \theta \)
  - Bernoulli: 
    \[
    P(\theta | D) = \frac{\theta^{\text{Count}(X=a)}(1-\theta)^{\text{Count}(X=\bar{a})}}{B(\text{Count}(X=a), \text{Count}(X=\bar{a}))} \sim \text{Beta}(\text{Count}(X=a) + \alpha_f, \text{Count}(X=\bar{a}) + \alpha_f)
    \]
  - Multinomial: 
    \[
    P(\theta | D) = \frac{\prod_{j=1}^{K} \theta_j^{\text{Count}(X=j)}}{D(\text{Count}(X_1), \ldots, \text{Count}(X_K))} \sim \text{Dirichlet}(\text{Count}(X_1) + \alpha_1, \ldots, \text{Count}(X_K) + \alpha_K)
    \]

- MAP estimate: 
  \[
  \hat{\theta} = \arg \max_{\theta} P(\theta | D) =
  \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!
MAP for the parameters of NB

- Dataset of $N$ examples
  - Let $\beta_{ib}=\text{Count}(X_i=a,Y=b)$ ← number of examples where $X_i=a$ and $Y=b$
  - Let $\gamma_i=\text{Count}(Y=b)$

- Prior
  \[
  Q(X|Y) \propto \text{Multinomial}(\alpha_{i1}, \ldots, \alpha_{iK}) \text{ or Multinomial}(\alpha/K) \\
  Q(Y) \propto \text{Multinomial}(\tau_{i1}, \ldots, \tau_{iM}) \text{ or Multinomial}(\tau/M)
  \]
  $m$ “virtual” examples

- MAP estimate
  \[
  \hat{\pi}_k = \arg \max_{\pi_k} \prod_k P(Y = y_k; \pi_k) P(\pi_k|\tau) = ? \\
  \hat{\theta}_{ijk} = \arg \max_{\theta_{ijk}} \prod_j P(X_i = x_{ij}|Y = y_k; \theta_{ijk}) P(\theta_{ijk}|\tau) = ?
  \]
  - Now, even if you never observe a feature/class, posterior probability never zero

What if we have continuous $X_i$?

- Eg., character recognition: $X_i$ is $i^{th}$ pixel

- Gaussian Naïve Bayes (GNB):
  \[
  P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}
  \]
  Sometimes assume variance
  - is independent of $Y$ (i.e., $\sigma_i$),
  - or independent of $X_i$ (i.e., $\sigma_k$)
  - or both (i.e., $\sigma$)
Estimating Parameters:  
Y discrete, X_i continuous

- Maximum likelihood estimates:

\[
\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j x_i^j \delta(Y^j = y_k),
\]

\[
\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_i (x_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k),
\]

\(\delta(x) = 1 \text{ if } x \text{ true, else } 0\)

Gaussian Naïve Bayes
Brain scans can track activation with precision and sensitivity

Example: GNB for classifying mental states

~1 mm resolution
~2 images per sec.
15,000 voxels/image
non-invasive, safe
measures Blood Oxygen Level Dependent (BOLD) response

Typical impulse response

[Mitchell et al.]
Gaussian Naïve Bayes: Learned $\mu_{\text{voxel, word}}$

$P(\text{BrainActivity} \mid \text{WordCategory} = \{\text{People, Animal}\})$

Pairwise classification accuracy: 85%
What you need to know about Naïve Bayes

- Optimal decision using Bayes Classifier
- Naïve Bayes classifier
  - What's the assumption
  - Why we use it
  - How do we learn it
  - Why is Bayesian estimation important
- Text classification
  - Bag of words model
- Gaussian NB
  - Features are still conditionally independent
  - Each feature has a Gaussian distribution given class