Classification

- **Learn**: $h : X \mapsto Y$
  - $X$ – features
  - $Y$ – target classes

- Suppose you know $P(Y|X)$ exactly, how should you classify?
  - Bayes classifier:
    $$y^* = \arg\max_y P(Y=y|X=x)$$

- Why?
**Optimal classification**

- **Theorem:** Bayes classifier $h_{\text{Bayes}}$ is optimal!
  
  $$h_{\text{Bayes}}(x) = \arg\max_y P(Y = y | X = x)$$

  □ That is $\text{error}_{\text{true}}(h_{\text{Bayes}}) \leq \text{error}_{\text{true}}(h)$, $\forall h(x)$

- **Proof:**
  
  $$p(\text{error}_h) = \int_x p(\text{error}_h | x) p(x) dx$$

  if $I$ say $h(x) = \text{kind}$

  $p(\text{error}_h | x) \ll \text{small}$

  $p(\text{error}_h | x)$

  $= p(Y = \text{not kind} | X)

  \text{as small as possible} \Rightarrow h_{\text{Bayes}}$

---

**Bayes Rule**

$$P(Y | X) = \frac{P(X | Y) P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$
How hard is it to learn the optimal classifier?

- Data =

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- How do we represent these? How many parameters?

  (not $k$, because $z$)

  - Prior, $P(Y)$:
    - Suppose $Y$ is composed of $k$ classes
      
      \[
      P(Y) \propto \prod_{i=1}^{k-1} P(Y_i) \]

  - Likelihood, $P(X|Y)$:
    - Suppose $X$ is composed of $n$ binary features
      
      \[
      P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | Y = y) = \left( \frac{2^n - 1}{k} \right)^{n \text{ settings}}
      \]

  - Complex model! High variance with limited data!!!

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) $

- e.g., $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$

- Thunder indep. of Rain given Lightning

- Not Thunder indep. Rain

- Equivalent to:

  \[
P(X, Y | Z) = P(X | Z)P(Y | Z)
  \]
What if features are independent?

- Predict Thunder
- From two **conditionally Independent** features
  - Lightening
  - Rain

Estimate $P(X|Y)$:

$P(L, R | T)$ \( \leq \) 6 params

Independently:

$P(L, R | T) = P(L | T) P(R | T)$

2: (2-1) \( \times \) 2 parameters

4 parameters

The Naïve Bayes assumption

- Naïve Bayes assumption:
  - Features are 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:
    
    $P(X_1...X_n | Y) = \prod_i P(X_i | Y)$

- How many parameters now?
  - Suppose $X$ is composed of $n$ binary features
    
    $P(X_1 | Y) \leq (2-1) k$

    for the entire conditional probability $P(X_1 | Y)$, assuming

    $\leq$ $n k$ parameters

    $\approx$ $2^k$ parameters

    instead of $2^n$
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!

MLE for the parameters of NB

- Given dataset
  - Count($A=a, B=b$) is number of examples where $A=a$ and $B=b$

- MLE for NB, simply:
  - Prior: $P(Y=y) = \frac{\text{count}(Y=y)}{m}$
  - Likelihood: $P(X=x_i | Y=y) = \frac{\text{count}(X_i=x_i, Y=y)}{\text{count}(Y=y)}$
Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:
  \[
P(X_1 \ldots X_n \mid Y) \neq \prod_i P(X_i \mid Y)
\]

- Actual probabilities \(P(Y \mid X)\) often biased towards 0 or 1
- Nonetheless, NB is the single most used classifier out there
  - NB often performs well, even when assumption is violated
  - [Domingos & Pazzani '96] discuss some conditions for good performance

Subtleties of NB classifier 2 – Insufficient training data

- What if you never see a training instance where \(X_1 = a\) when \(Y = b\)?
  - e.g., \(Y = \{\text{SpamEmail}\}, X_1 = \{\text{‘Enlargement’}\}
  - \(P(X_1 = a \mid Y = b) = 0\)
  - \(P(\{\text{Enlargement}\} = 0 = \frac{\text{Count}(X_1 = \text{‘Enlargement’}, Y = \text{SpamEmail})}{\text{Count}(Y = \text{SpamEmail})}\)

- Thus, no matter what the values \(X_2, \ldots, X_n\) take:
  - \(P(Y = b \mid X_1 = a, X_2, \ldots, X_n) = 0\)

- What now???
**MAP for Beta distribution**

\[ P(\theta \mid D) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- MAP: use most likely parameter:
  \[ \hat{\theta} = \arg \max_{\theta} P(\theta \mid D) = \frac{\beta_H + \alpha_H - 1}{\beta_H + \alpha_H + \beta_T + \alpha_T - 2} \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \to 1 \), prior is “forgotten”
- **But, for small sample size, prior is important!**

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

- Dataset of \( N \) examples
- Prior
  - “distribution” \( Q(X_i, Y) \), \( Q(Y) \)
  - \( m \) “virtual” examples
- MAP estimate
  - \( P(X \mid Y) \)

Now, even if you never observe a feature/class, posterior probability never zero
Text classification

- Classify e-mails
  - $Y = \{\text{Spam, NotSpam}\}$
- 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 $i^{\text{th}}$ word in article

```
Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
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
Zhitnuk. 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 Hruday 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!!!
  - $P(X=x|Y=y)$ is just the probability of observing word $x_i$ in a document on topic $y$

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

Bag of words model

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

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

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}^{Length Doc} P(x_i | y) \rightarrow \hat{P}(x_i = x_i | Y = y) \approx \frac{\text{Count } ("\text{word } \text{in span}\)}}{\text{Count } (\text{of words in span})}
\]

Bag of Words Approach

- **Aardvark** 0
- **About** 2
- **All** 2
- **Africa** 1
- **Apple** 0
- **Anxious** 0
- **..., Gas** 1
- **..., Oil** 1
- **..., Zaire** 0

**TOTAL**

- Our energy exploration, production, and distribution operations span the globe, with operations in more than 100 countries.
- As TOTAL, we draw on our greatest strength from our fast-growing oil and gas sectors. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.
- Our expanding refining and marketing operations in Asia and the Mediterranean Sea complement already solid positions in Europe, Africa, and the U.S.
- Our growing specialty chemicals sector adds balance and profit to the core energy business.
NB with Bag of Words for text classification

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
  - Use naïve Bayes decision rule

\[ h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i|y) \]

Twenty News Groups results

Given 1000 training documents from each group Learn to classify new documents according to which newsgroup it came from

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

Naive Bayes: 89% classification accuracy
Learning curve for Twenty News Groups

Accuracy vs. Training set size (1/3 withheld for test)