# Naïve Bayes

The University of Texas at Dallas

# Supervised Learning of Classifiers

## Find $f$

- **Given:** Training set $\{(x_i, y_i) \mid i = 1 \ldots n\}$
- **Find:** A good approximation to $f : X \rightarrow Y$

**Examples:** what are $X$ and $Y$ ?

- Spam Detection
  - Map email to {Spam,Ham}
- Digit recognition
  - Map pixels to {0,1,2,3,4,5,6,7,8,9}
- Stock Prediction
  - Map new, historic prices, etc. to Â (the real numbers)

Classification

# Bayesian Categorization/Classification

- Let the set of categories be $\{c_1, c_2, \ldots c_n\}$
- Let $E$ be description of an instance.
- Determine category of $E$ by determining for each $c_i$

$$P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}$$

- $P(E)$ can be ignored (normalization constant)

$$P(c_i \mid E) \sim P(c_i)P(E \mid c_i)$$

- Select the class with the max. probability.

# Text classification

- Classify e-mails
  - Y = {Spam,NotSpam}
- Classify news articles
  - Y = {what is the topic of the article?}
- Classify webpages
  - Y = {Student, professor, project, …}

- What to use for features, **X**?

# Features **X** are word sequence in document
# $X_i$ for $i^{th}$ 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
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
```

# Features for Text Classification

- **X** is sequence of words in document
- **X** (and hence 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.
- $10{,}000^{1000} = 10^{4000}$
- Atoms in Universe: $10^{80}$
  - We may have a problem…

# Bag of Words Model

Typical additional assumption –

- **Position in document doesn't matter**:
  - $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - (all positions have the same distribution)
- Ignore the order of words
- Sounds really silly, but often works very well!

- Features
  - **X** = Set of all possible words
  - Value of the variable = Frequency (number of times it occurs) in the document

# Bag of Words Approach

the world of TOTAL

All About The Company
Global Activities
Corporate Structure
TOTAL's Story
Upstream Strategy
Downstream Strategy
Chemicals Strategy
TOTAL Foundation
Homepage

*all about the* **company**

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

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. 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 Rim 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.

| | |
|---|---|
| aardvark | 0 |
| about | 2 |
| all | 2 |
| Africa | 1 |
| apple | 0 |
| anxious | 0 |
| ... | |
| gas | 1 |
| ... | |
| oil | 1 |
| … | |
| Zaire | 0 |

# Bayesian Categorization

$P(y_1 \mid \mathbf{X}) \sim P(y_i)P(\mathbf{X} \mid y_i)$

Problem!

- Need to know:
  - Priors: $P(y_i)$
  - Conditionals: $P(\mathbf{X} \mid y_i)$
- $P(y_i)$ are easily estimated from data.
  - If $n_i$ of the examples in $D$ are in $y_i$, then $P(y_i) = n_i / |D|$
- Conditionals:
  - $\mathbf{X} = X_1 \wedge \ldots \wedge X_n$
  - Estimate $P(X_1 \wedge \ldots \wedge X_n \mid y_i)$
- Too many possible instances to estimate!
  - *(exponential in n)*
  - Even ***with*** bag of words assumption!

# Need to Simplify Somehow

- Too many probabilities
  - $P(x_1 \wedge x_2 \wedge x_3 \mid y_i)$

$P(x_1 \wedge x_2 \wedge x_3 \mid spam)$
$P(x_1 \wedge x_2 \wedge \neg x_3 \mid spam)$
$P(x_1 \wedge \neg x_2 \wedge x_3 \mid spam)$
....
$P(\neg x_1 \wedge \neg x_2 \wedge \neg x_3 \mid \neg spam)$

- Can we assume some are the same?
  - $P(x_1 \wedge x_2 | y_i) = P(x_1 | y_i)\, P(x_2 | y_i)$

# Conditional Independence

- X is **conditionally independent** of Y given Z, if the probability distribution for 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.,

$$P(Thunder | Rain, Lightning) = P(Thunder | Lightning)$$

- Equivalent to:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

# Naïve Bayes

- Naïve Bayes assumption:
  - Features are independent given class:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

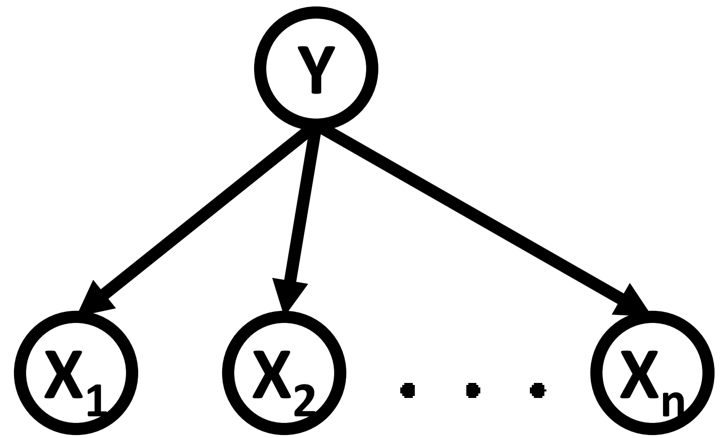
  - More generally:

$$\boxed{P(X_1 ... X_n|Y) = \prod_i P(X_i|Y)}$$

- How many parameters now?
  - Suppose **X** is composed of *n* binary features

# The Naïve Bayes Classifier

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




## Decision rule:

$$y^* = h_{NB}(\mathbf{x}) = \arg\max_y P(y)P(x_1, \ldots, x_n \mid y)$$
$$= \arg\max_y P(y) \prod_i P(x_i|y)$$

# MLE for the parameters of NB

- Given dataset, count occurrences for all pairs
  - $Count(X_i = x, Y = y)$
  - How many pairs?
- MLE for discrete NB, simply:
  - Prior:

$$P(Y = y) = \frac{Count(Y = y)}{\sum_{y'} Count(Y = y')}$$

  - Likelihood:

$$P(X_i = x | Y = y) = \frac{Count(X_i = x, Y = y)}{\sum_{x'} Count(X_i = x', Y = y)}$$

# NAÏVE BAYES CALCULATIONS

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

# Subtleties of NB Classifier: #1 Violating the NB Assumption

- Usually, features are not conditionally independent:

$$\boxed{P(X_1 \ldots X_n | Y) \neq \prod_i P(X_i | Y)}$$

- The naïve Bayes assumption is often violated, yet it performs surprisingly well in many cases.
- Plausible reason: Only need the probability of the correct class to be the largest!
  - Example: two-way classification; just need to figure out the correct side of 0.5 and not the actual probability (0.51 is the same as 0.99).

# Subtleties of NB Classifier: #2 Insufficient Training Data

- What if you never see a training instance $(X_1 = a, Y = b)$
  - Example: you did not see the word Enlargment in spam!
  - Then $\Pr(X_1 = a | Y = b) = 0$
- Thus no matter what values $X_2, \cdots, X_n$ take:
  - $P(X_1 = \text{Enlargment}, X_2 = a, \cdots, X_n = a | Y = b) = 0$
  - Why?

$$
\begin{aligned}
y^* = h_{NB}(\mathbf{x}) &= \arg\max_y P(y) P(x_1, \ldots, x_n \mid y) \\
&= \arg\max_y P(y) \prod_i P(x_i | y)
\end{aligned}
$$

# For Binary Features: We already know the answer!

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

- MAP: use most likely parameter

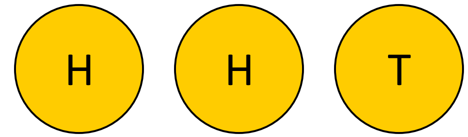
$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H}{\alpha_H + \beta_H + \alpha_T + \beta_T}$$

- Beta prior equivalent to extra observations for each feature
- As $N \rightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

# Multinomials: Laplace Smoothing

- Laplace's estimate:
  - Pretend you saw every outcome k extra times

$$P_{LAP,k}(x) = \frac{c(x)+k}{N+k|X|}$$




$$P_{LAP,0}(X) = \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$$

$$P_{LAP,1}(X) = \left\langle \frac{3}{5}, \frac{2}{5} \right\rangle$$

$$P_{LAP,100}(X) = \left\langle \frac{102}{203}, \frac{101}{203} \right\rangle$$

  - What's Laplace with k = 0?
  - k is the strength of the prior
  - Can derive this as a MAP estimate for multinomial with *Dirichlet priors*

- Laplace for conditionals:
  - Smooth each condition independently:

$$P_{LAP,k}(x|y) = \frac{c(x,y)+k}{c(y)+k|X|}$$

# Probabilities: Important Detail!

- $P(\text{spam} \mid X_1 \ldots X_n) = \prod_i P(\text{spam} \mid X_i)$

Any more potential problems here?

- We are multiplying lots of small numbers
  Danger of underflow!
  - $0.5^{57} = 7\text{ E }-18$
- Solution? Use logs and add!
  - $p_1 * p_2 = e^{\log(p1)+\log(p2)}$
  - Always keep in log form

# Naïve Bayes: Summary

Model: Given a set of $n$ features, denoted by $\mathbf{X}$ and a class variable $Y$

$$P(\mathbf{X}, Y) = P(Y) \prod_{i=1}^{n} P(X_i | Y)$$

Learning Task: Given a dataset $\mathcal{D}$, estimate $P(Y)$; $P(X_i | Y)$

Learning Algorithm:

$$P(Y = y) = \frac{Count_{\mathcal{D}}(Y = y)}{|\mathcal{D}|}$$

$$P(X_i = x_i | Y = y) = \frac{Count_{\mathcal{D}}(X_i = x_i, Y = y) + K}{Count_{\mathcal{D}}(Y = y) + K|X_i|}$$

# Naïve Bayes: Summary

Classification: Given a test example $(X_1 = x_1, \ldots, X_n = x_n)$, compute the following quantity for each class $Y = y$ and choose the class with the maximum value

$$P(Y = y) \prod_{i=1}^{n} P(X_i = x_i | Y = y)$$

In practice, store in log-space, compute the following quantity and choose the class having the maximum value:

$$w(Y = y) \sum_{i=1}^{n} w_i(X_i = x_i | Y = y)$$

where $w(Y = y) = \log(P(Y = y))$ and
$w_i(X_i = x_i | Y = y) = \log(P(X_i = x_i | Y = y))$

# NB for Text Classification: Learning

- Learning phase: $P(Y_m)$ and $P(X_i|Y_m)$

Prior: $P(Y_m)$

$$P(Y_m) = \frac{N_m}{N}$$

where $N_m$ is the number of documents having class label $m$ and $N$ is the total number of documents.

Class conditional probabilities: $P(X_i|Y_m)$

$$P(X_i|Y_m) = \frac{Count(X_i, Y_m) + 1}{\sum_{j=1}^{V}(Count(X_j, Y_m) + 1)}$$

where $V$ is the size of the vocabulary (number of distinct words) in all documents and $Count(X_i, Y_m)$ is the number of times the word $X_i$ appears in documents of class $Y_m$.

# NB for Text Classification: Classification

- Given a new document having length "L"

$$\arg\max_{Y} P(Y) \prod_{i=1}^{L} P(X_i | Y)$$

# Example: (Borrowed from Dan Jurafsky)

| | Doc | Words | Class |
|---|---|---|---|
| Training | 1 | Chinese Beijing Chinese | c |
| | 2 | Chinese Chinese Shanghai | c |
| | 3 | Chinese Macao | c |
| | 4 | Tokyo Japan Chinese | j |
| Test | 5 | Chinese Chinese Chinese Tokyo Japan | ? |

**Priors:**

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

**Conditional Probabilities:**

$P(\text{Chinese}|c) = (5+1) / (8+6) = 6/14 = 3/7$

$P(\text{Tokyo}|c) = (0+1) / (8+6) = 1/14$

$P(\text{Japan}|c) = (0+1) / (8+6) = 1/14$

$P(\text{Chinese}|j) = (1+1) / (3+6) = 2/9$

$P(\text{Tokyo}|j) = (1+1) / (3+6) = 2/9$

$P(\text{Japan}|j) = (1+1) / (3+6) = 2/9$

**Choosing a class:**

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

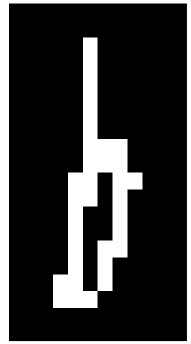
$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

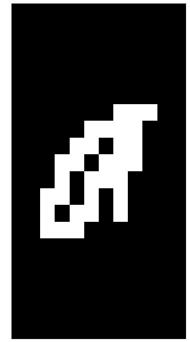
# Bayesian Learning

# What if Features are Continuous?

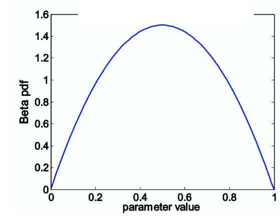
Eg., Character Recognition:

$X_i$ is i[th] pixel

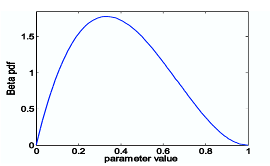




Prior




Posterior




$$P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\ P(Y)$$

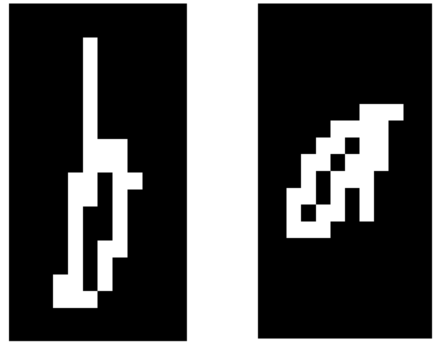
Data Likelihood

## Bayesian Learning

# What if Features are Continuous?

Eg., Character Recognition:

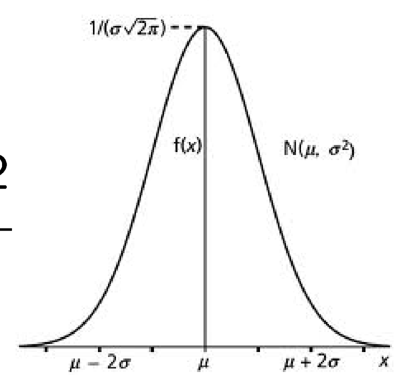
$X_i$ is $i^{th}$ pixel



$$P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\, P(Y)$$

$$P(X_i = x \mid Y = y_k) = N(\mu_{ik}, \sigma_{ik})$$

$$N(\mu_{ik}, \sigma_{ik}) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$



# Gaussian Naïve Bayes

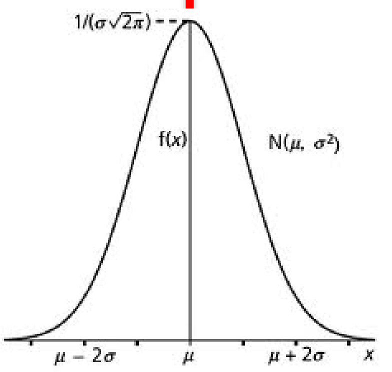
**Sometimes Assume Variance**

- is independent of Y (i.e., $\sigma_i$),
- or independent of $X_i$ (i.e., $\sigma_k$)
- or both (i.e., $\sigma$)

$$P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\, P(Y)$$

$$P(X_i = x \mid Y = y_k) = N(\mu_{ik}, \sigma_{ik})$$

$$N(\mu_{ik}, \sigma_{ik}) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$



# Learning Gaussian Parameters

## Maximum Likelihood Estimates:

- Mean:

$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Variance:

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2$$

# Learning Gaussian Parameters

## Maximum Likelihood Estimates:

- Mean:

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

j^th training example

δ(x)=1 if x true, else 0

- Variance:

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

# Learning Gaussian Parameters

## Maximum Likelihood Estimates:

- Mean:

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

- Variance:

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

# What you need to know about Naïve Bayes

- 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