Learning Distribution (Parametric Approach)

Dr. Jianlin Cheng

Electrical Engineering and Computer Science Department
University of Missouri, Columbia
Fall, 2019

Slides Adapted from Book, CMU and Stanford Machine Learning Courses, my research
Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what’s the probability it will fall with the head up?
  - You say: Please flip it a few times:
Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what’s the probability it will fall with the head up?
  - You say: Please flip it a few times:

- Your answer: 3 / 5
- He says: Why?
- You say: Because
Bernoulli distribution

Data, $D = \frac{\text{num. of heads}}{\text{total observations}}$

• $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1-\theta$

• Flips are i.i.d.:
  – Independent events
  – Identically distributed according to Bernoulli distribution

Choose $\theta$ that maximizes the probability of observed data
Maximum Likelihood Estimation

Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D \mid \theta)$$
Maximum Likelihood Estimation

Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D | \theta)$$

$$P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = P(H) \times P(T) \times P(H) \times P(H) \times P(T) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

$$\log P(D|\theta) = \alpha_H \log \theta + \alpha_T \log (1 - \theta)$$

$$\frac{d \log P(D|\theta)}{d\theta} = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0$$

$$\alpha_H - (\alpha_H + \alpha_T)\theta = 0$$
Maximum Likelihood Estimation

Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_\theta P(D | \theta)$$

MLE of probability of head:

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{3}{5}$$

“Frequency of heads”
How many flips do I need?

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: \( \theta = \frac{3}{5} \), I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What’s better?**
- You say: Hmm... The more the merrier???
- **He says: Is this why I am paying you the big bucks???
Simple bound (Hoeffding’s inequality)

- For $n = \alpha_H + \alpha_T$, and $\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

- Let $\theta^*$ be the true parameter, for any $\varepsilon > 0$:

$$P(|\hat{\theta} - \theta^*| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$$
Homework assignment 1: derive the bound for $n$ given $\epsilon$, $\delta$

PAC Learning

• PAC: Probably Approximately Correct
• Billionaire says: I want to know the coin parameter $\theta$, within $\epsilon = 0.1$, with probability at least $1-\delta = 0.95$. How many flips?

$$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Sample complexity

$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Homework assignment 1: derive the bound for $n$ given $\epsilon$, $\delta$
What about prior knowledge?

- Billionaire says: Wait, I know that the coin is “close” to 50-50. What can you do for me now?
- You say: I can learn it the Bayesian way...

- Rather than estimating a single $\theta$, we obtain a distribution over possible values of $\theta$. 

![Diagram showing the transition from prior to posterior distribution](image)
Bayesian Learning

- Use Bayes rule:

\[ P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \]

- Or equivalently:

\[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]

posterior likelihood prior

Prior distribution

• What about prior?
  – Represents expert knowledge (philosophical approach)
  – Simple posterior form (engineer’s approach)

• Uninformative priors:
  – Uniform distribution

• Conjugate priors:
  – Closed-form representation of posterior
  – $P(\theta)$ and $P(\theta | D)$ have the same form
Conjugate Prior

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

**Eg. 1 Coin flip problem**

Likelihood is $\sim$ Binomial

$$P(D | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H-1} (1 - \theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$
Beta distribution

$\text{Beta}(\beta_H, \beta_T)$

More concentrated as values of $\beta_H$, $\beta_T$ increase
Beta conjugate prior

\[ P(\theta) \sim \text{Beta}(\beta_H, \beta_T) \]

\[ P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

As \( n = \alpha_H + \alpha_T \) increases

As we get more samples, effect of prior is “washed out”
Conjugate Prior

- $P(\theta)$ 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.
Maximum A Posteriori Estimation

Choose $\theta$ that maximizes a posterior probability

$$\hat{\theta}_{MAP} = \arg \max_\theta P(\theta | D)$$

$$= \arg \max_\theta P(D | \theta)P(\theta)$$

MAP estimate of probability of head:

$$P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution
MLE vs. MAP

- Maximum Likelihood estimation (MLE)
  Choose value that maximizes the probability of observed data
  \[
  \hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)
  \]

- Maximum a posteriori (MAP) estimation
  Choose value that is most probable given observed data and prior belief
  \[
  \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta|D) = \arg \max_{\theta} P(D|\theta)P(\theta)
  \]

When is MAP same as MLE?
MAP using Conjugate Prior

\[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta \mid D) = \arg \max_{\theta} P(D \mid \theta)P(\theta) \]

Coin flip problem

Likelihood is \( \sim \) Binomial

\[ P(D \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T} \]

If prior is Beta distribution,

\[ P(\theta) \propto \theta^{\beta_H - 1}(1 - \theta)^{\beta_T - 1} \sim Beta(\beta_H, \beta_T) \]

Then posterior is Beta distribution

\[ P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

For Binomial, conjugate prior prior is Beta distribution.
A few beta probability distributions

- $B(10,10)$
- $B(1,1)$
- $B(2,2)$
- $B(5,2)$
- $B(1,5)$
MLE vs. MAP

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

What if we toss the coin too few times?

• You say: Probability next toss is a head = 0
• Billionaire says: You’re fired! ...with prob 1 😊

\[ \hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

• Beta prior equivalent to extra coin flips (regularization)
• As \( n \to \infty \), prior is “forgotten”
• But, for small sample size, prior is important!
Bayesians vs. Frequentists

You are no good when sample is small

You give a different answer for different priors
What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- You say: Let me tell you about Gaussians...

\[
P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = N(\mu, \sigma^2)
\]
Gaussian distribution

Data, $D =$

$$
\begin{array}{cccccccc}
3 & 4 & 5 & 6 & 7 & 8 & 9 & \text{Sleep hrs} \\
\end{array}
$$

- Parameters: $\mu$ – mean, $\sigma^2$ – variance
- Sleep hrs are i.i.d.:
  - Independent events
  - Identically distributed according to Gaussian distribution
Properties of Gaussians

- **Affine transformation** (multiplying by scalar and adding a constant)
  - $X \sim N(\mu, \sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b, a^2\sigma^2)$

- **Sum of Gaussians**
  - $X \sim N(\mu_X, \sigma_X^2)$
  - $Y \sim N(\mu_Y, \sigma_Y^2)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$
MLE Estimate of Gaussian

• Find \( u, \sigma^2 \) that maximize \( P(D|u, \sigma^2) \)
MLE Estimate of Gaussian

- Find $u, \sigma^2$ that maximize $P(D|u, \sigma^2)$

$$P(D|u, \sigma^2) = \prod_{i=1}^{n} P(x_i|u, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-u)^2}{2\sigma^2}}$$

$$\log P(D|u, \sigma^2) = \sum_{i=1}^{n} \left( -\log(\sqrt{2\pi\sigma}) - \frac{(x_i-u)^2}{2\sigma^2} \right) = -n\log(\sqrt{2\pi\sigma}) - \sum_{i=1}^{n} \frac{(x_i-u)^2}{2\sigma^2}$$

$$\frac{\partial \log P(D|u,\sigma^2)}{\partial u} = -\sum_{i=1}^{n} \frac{-2(x_i-u)}{2\sigma^2} = +\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} x_i - nu = 0$$
MLE for Gaussian mean and variance

\[ \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ \hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]

Note: MLE for the variance of a Gaussian is **biased**
- Expected result of estimation is **not** true parameter!
- Unbiased variance estimator:
\[ \hat{\sigma}^2_{unbiased} = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]
MAP for Gaussian mean and variance

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

\[
P(\mu \mid \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}} = \mathcal{N}(\eta, \lambda^2)
\]
MAP for Gaussian Mean

\[ \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ \hat{\mu}_{MAP} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2} \]

(Assuming known variance \(\sigma^2\))
What you should know...

• Learning parametric distributions: form known, parameters unknown
  – Bernoulli ($\theta$, probability of flip)
  – Gaussian ($\mu$, mean and $\sigma^2$, variance)
• MLE
• MAP
What loss function are we minimizing?

- Learning distributions/densities

- **Task:** Learn $P(X; \theta) \equiv \text{Learn } \theta$ (know form of $P$, except $\theta$)

- **Experience:** $D = \{X_i\}_{i=1}^{n} \sim P(X; \theta)$

- **Performance:**
  \[
  \max_{\theta} P(D|\theta) \\
  = \min_{\theta} - \log P(D|\theta) \\
  = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} - \log P(X_i|\theta)
  \]

  **Negative log Likelihood loss**
Learn a Probabilistic Classifier

**Task:** Predict whether or not a picnic spot is enjoyable

**Training Data:**

\[ X = (X_1, X_2, X_3, \ldots, X_d) \]

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySpt</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>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Let's learn** \( P(Y|X) \) – how many parameters?

Prior: \( P(Y = y) \) for all \( y \) \( \text{K-1 if K labels} \)

Likelihood: \( P(X=x|Y = y) \) for all \( x, y \) \( (2^d - 1)K \) if \( d \) binary features
Learning the Optimal Classifier

Task: Predict whether or not a picnic spot is enjoyable

Training Data: \( X = (X_1, X_2, X_3, \ldots, X_d) \)

\[
\begin{array}{cccccccc}
\text{Sky} & \text{Temp} & \text{Humid} & \text{Wind} & \text{Water} & \text{Forecast} & \text{EnjoySpt} \\
\text{Sunny} & \text{Warm} & \text{Normal} & \text{Strong} & \text{Warm} & \text{Same} & \text{Yes} \\
\text{Sunny} & \text{Warm} & \text{High} & \text{Strong} & \text{Warm} & \text{Same} & \text{Yes} \\
\text{Rainy} & \text{Cold} & \text{High} & \text{Strong} & \text{Warm} & \text{Change} & \text{No} \\
\text{Sunny} & \text{Warm} & \text{High} & \text{Strong} & \text{Cool} & \text{Change} & \text{Yes} \\
\end{array}
\]

\( n \) rows

Let's learn \( P(Y|X) \) — how many parameters?

\[ 2^dK - 1 \] (\( K \) classes, \( d \) binary features)

Need \( n >> 2^dK - 1 \) number of training data to learn all parameters
Conditional Independence

- **X** is conditionally independent of **Y** given **Z**: The probability distribution governing **X** is independent of the value of **Y**, given the value of **Z**.

  $$(\forall x, y, z) P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

- Equivalent to:

  $$P(X, Y | Z) = P(X | Z)P(Y | Z)$$

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

**Note:** does NOT mean Thunder is independent of Rain.
Conditional vs. Marginal Independence

- C calls A and B separately and tells them a number $n \in \{1, \ldots, 10\}$
- Due to noise in the phone, A and B each imperfectly (and independently) draw a conclusion about what the number was.
- A thinks the number was $n_a$ and B thinks it was $n_b$.
- Are $n_a$ and $n_b$ marginally independent?
  - No, we expect e.g. $P(n_a = 1 \mid n_b = 1) > P(n_a = 1)$
- Are $n_a$ and $n_b$ conditionally independent given $n$?
  - Yes, because if we know the true number, the outcomes $n_a$ and $n_b$ are purely determined by the noise in each phone.
    
    $P(n_a = 1 \mid n_b = 1, n = 2) = P(n_a = 1 \mid n = 2)$
Prediction using Conditional Independence

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

# parameters needed to learn likelihood given L

\[ P(T,R|L) = (2^2-1)2 = 6 \]

With conditional independence assumption

\[ P(T,R|L) = P(T|L) P(R|L) \]

\[ (2-1)2 + (2-1)2 = 4 \]
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:
    \[ P(X_1...X_d|Y) = \prod_{i=1}^{d} P(X_i|Y) \]

- How many parameters now?
  - Suppose \( X \) is composed of \( d \) binary features
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:
    \[ P(X_1 \ldots X_d|Y) = \prod_{i=1}^{d} P(X_i|Y) \]
- How many parameters now? \((2-1)dK\) vs. \((2^d-1)K\)
  - Suppose \(X\) is composed of \(d\) binary features
Naïve Bayes Classifier

• Given:
  – Class Prior $P(Y)$
  – $d$ conditionally independent features $X$ given the class $Y$
  – For each $X_i$, we have likelihood $P(X_i | Y)$

• Decision rule:

$$f_{NB}(x) = \arg \max_y P(x_1, \ldots, x_d | y)P(y)$$

$$= \arg \max_y \prod_{i=1}^{d} P(x_i | y)P(y)$$

• If conditional independence assumption holds, NB is optimal classifier! But worse otherwise.
Naïve Bayes Algo – Discrete features

- **Training Data** \( \{(X^{(j)}, Y^{(j)})\}_{j=1}^{n} \) \( X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)}) \)

- **Maximum Likelihood Estimates**
  - For Class Prior
    \[
    \hat{P}(y) = \frac{\#j: Y^{(j)} = y}{n}
    \]
  - For Likelihood
    \[
    \hat{P} = (x_i | y) = \frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\#j: X_i^{(j)} = x_i, Y^{(j)} = y}{\#j: Y^{(j)} = y}/n
    \]

- **NB Prediction for test data** \( X = (x_1, \ldots, x_d) \)

\[
Y = \arg \max_y \hat{P}(y) \prod_{i=1}^{d} \frac{\hat{P}(x_i, y)}{\hat{P}(y)}
\]
Subtlety 1 – Violation of NB Assumption

- Usually, features are not conditionally independent:

\[ P(X_1 \ldots X_d | Y) \neq \prod_i P(X_i | Y) \]

- 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
Subtlety 2 – Insufficient training data

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

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

$$P(X_1 = a, X_2, ..., X_n \mid Y) = P(X_1 = a \mid Y) \prod_{i=2}^{d} P(X_i \mid Y)$$

- What now???
MLE vs. MAP

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You’re fired! ...with prob 1 😊

\[ \hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

- Beta prior equivalent to extra coin flips
- As \( N \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!
Naïve Bayes Algo – Discrete features

- Training Data \( \{(X^{(j)}, Y^{(j)})\}_{j=1}^{n} \) \( X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)}) \)

- Maximum A Posteriori Estimates – add \( m \) “virtual” examples

  Assume priors

  \[ Q(Y = b) \quad Q(X_i = a, Y = b) \]

  MAP Estimate

  \[ \hat{P}(X_i = a | Y = b) = \frac{\{ \# j : X_i^{(j)} = a, Y^{(j)} = b \}}{\{ \# j : Y^{(j)} = b \}} + \frac{m Q(X_i = a, Y = b)}{\# \text{ virtual examples with } Y = b} \]

Now, even if you never observe a class/feature posterior probability never zero.
Case Study: 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!
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
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!!!
  - \( P(X_i=x_i|Y=y) \) is just the probability of observing word \( x_i \) at the \( i^{th} \) position 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!

\[
\prod_{i=1}^{\text{LengthDoc}} P(x_i | y) = \prod_{w=1}^{W} P(w | y)^{\text{count}_w}
\]
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.
Twenty news groups results

<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>
Learning curve for twenty news groups

Accuracy vs. Training set size (1/3 withheld for test)
What if features are continuous?

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

Gaussian Naïve Bayes (GNB):

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Different mean and variance for each class $k$ and each pixel $i$.

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:

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

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^{N} x_j$$

$$\hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{j=1}^{N} (x_j - \hat{\mu})^2$$

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

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

[Mitchell et al.]

15,000 voxels or features

10 training examples or subjects per class
Learned Naïve Bayes Models – Means for $P(\text{BrainActivity} \mid \text{WordCategory})$

Pairwise classification accuracy: 85%  
[Mitchell et al.]
What you should know...

- 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