Lecture 8:
- Maximum Likelihood Estimation (MLE) (cont’d.)
- Maximum a posteriori (MAP) estimation
- Naïve Bayes Classifier
Last time... Flipping a Coin

I have a coin, if I flip it, what’s the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:

The estimated probability is: $\frac{3}{5}$ “Frequency of heads”
Last time... Flipping a Coin

The estimated probability is: $\frac{3}{5}$ “Frequency of heads”

Questions:
(1) Why frequency of heads???
(2) How good is this estimation???
(3) Why is this a machine learning problem???

We are going to answer these questions
Question (1)

Why frequency of heads???

• Frequency of heads is exactly the maximum likelihood estimator for this problem

• MLE has nice properties (interpretation, statistical guarantees, simple)
MLE for Bernoulli distribution

Data, \( D = \{X_i\}_{i=1}^{n}, \ X_i \in \{H, T\} \)

\[ P(\text{Heads}) = \theta, \ P(\text{Tails}) = 1-\theta \]
MLE for Bernoulli distribution

Data, \( D = \{X_i\}_{i=1}^n, \ X_i \in \{\text{H, T}\} \)

\[ P(\text{Heads}) = \theta, \ P(\text{Tails}) = 1 - \theta \]

Flips are i.i.d.:
MLE for Bernoulli distribution

Data, $D =$

$$D = \{X_i\}_{i=1}^{n}, \quad X_i \in \{H, T\}$$

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

Flips are i.i.d.:
- Independent events
- Identically distributed according to Bernoulli distribution
MLE for Bernoulli distribution

Data, \( D = \{X_i\}_{i=1}^n, \ X_i \in \{H, T\} \)

\[ P(\text{Heads}) = \theta, \ P(\text{Tails}) = 1-\theta \]

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

MLE: Choose \( \theta \) that maximizes the probability of observed data
Maximum Likelihood Estimation

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

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$
MLE: Choose $\theta$ that maximizes the probability of observed data

\[
\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta) = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)
\]

independent draws
Maximum Likelihood Estimation

**MLE:** Choose $\theta$ that maximizes the probability of observed data

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

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

$$
= \arg \max_\theta \prod_{i: X_i = H} \theta \prod_{i: X_i = T} (1 - \theta)
$$

*independent draws*

*identically distributed*
Maximum Likelihood Estimation

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

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

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

$$= \arg \max_\theta \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta)$$

$$= \arg \max_\theta \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

slide by Barnabás Póczos & Alex Smola
Maximum Likelihood Estimation

MLE: Choose \( \theta \) that maximizes the probability of observed data

\[
\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)
= \arg \max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)
= \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta)
= \arg \max_{\theta} \theta^\alpha_H (1 - \theta)^\alpha_T
\]

\[J(\theta)\]
**Maximum Likelihood Estimation**

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

\[
\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta) \\
= \arg \max_{\theta} \theta^{\alpha_H} (1 - \theta)^{\alpha_T} J(\theta)
\]
Maximum Likelihood Estimation

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

$$
\hat{\theta}_{MLE} = \operatorname{arg\,max}_\theta P(D|\theta) = \operatorname{arg\,max}_\theta \theta^{\alpha_H} (1 - \theta)^{\alpha_T} J(\theta)
$$

$$
\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1} (1 - \theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H} (1 - \theta)^{\alpha_T - 1} \bigg|_{\theta = \hat{\theta}_{MLE}} = 0
$$
Maximum Likelihood Estimation

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

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

$$= \arg \max_{\theta} \frac{\theta^{\alpha_H} (1 - \theta)^{\alpha_T}}{J(\theta)}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H-1} (1 - \theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H} (1 - \theta)^{\alpha_T-1} \bigg|_{\theta=\hat{\theta}_{MLE}} = 0$$

$$\alpha_H (1 - \theta) - \alpha_T \theta \bigg|_{\theta=\hat{\theta}_{MLE}} = 0$$
Question (2)

• How good is this MLE estimation???

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]
How many flips do I need?

I flipped the coins 5 times: 3 heads, 2 tails

\[ \hat{\theta}_{MLE} = \frac{3}{5} \]

What if I flipped 30 heads and 20 tails?

\[ \hat{\theta}_{MLE} = \frac{30}{50} \]

- Which estimator should we trust more?
- The more the merrier???
Simple bound

Let $\theta^*$ be the true parameter.

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

For any $\epsilon > 0$:

**Hoeffding’s inequality:**

$$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$
Probably Approximate Correct (PAC) Learning

I want to know the coin parameter $\theta$, within $\varepsilon = 0.1$ error with probability at least $1-\delta = 0.95$.

How many flips do I need?

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

Sample complexity:

$$n \geq \frac{\ln(2/\delta)}{2\varepsilon^2}$$
Question (3)

Why is this a machine learning problem???

- improve their performance (accuracy of the predicted prob.)
- at some task (predicting the probability of heads)
- with experience (the more coins we flip the better we are)
What about continuous features?

Let us try Gaussians...

\[ p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) = \mathcal{N}_x(\mu, \sigma) \]
MLE for Gaussian mean and variance

Choose $\theta= (\mu, \sigma^2)$ that maximizes the probability of observed data

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

$$= \arg \max_{\theta} \prod_{i=1}^{n} P(X_i \mid \theta)$$

Independent draws

$$= \arg \max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

Identically distributed

$$= \arg \max_{\theta= (\mu, \sigma^2)} \frac{1}{2\sigma^2} e^{-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}}$$

$$J(\theta)$$
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 the true parameter!]

Unbiased variance estimator:
\[ \hat{\sigma}^2_{unbiased} = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]
What about prior knowledge?
(MAP Estimation)
What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way...

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

Before data

After data

$P(\theta|D)$
Prior distribution

What prior? What distribution do we want for a 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
In order to proceed we will need:

Bayes Rule

Chain rule:

\[ P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \]

Bayes rule:

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

Bayes rule is important for reverse conditioning.
Bayesian Learning

- Use Bayes rule:
  \[ P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \]

- Or equivalently:
  \[ P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \]

  
  posterior  likelihood  prior
MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial

\[ P(\mathcal{D} | \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \]

If the prior is Beta distribution,

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

⇒ posterior is Beta distribution

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

P(\theta) and P(\theta | D) have the same form! [Conjugate prior]

\[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} P(D | \theta)P(\theta) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]
Beta distribution

More concentrated as values of $\alpha$, $\beta$ increase
Beta conjugate prior

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

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

As we get more samples, effect of prior is “washed out”
C3PO: Sir, the possibility of successfully navigating an asteroid field is approximately 3,720 to 1!

Han: Never tell me the odds!

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

https://www.countbayesie.com/blog/2015/2/18/hans-solo-and-bayesian-priors
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?**
From Binomial to Multinomial

Example: Dice roll problem (6 outcomes instead of 2)

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

\[
P(D | \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \cdots \theta_k^{\alpha_k}
\]

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.

Bayesians vs. Frequentists

You are no good when sample is small

You give a different answer for different priors
Recap: What about prior knowledge? (MAP Estimation)
Recap: What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way...

Rather than estimating a single $\theta$, we obtain a distribution over possible values of $\theta$.
Recap: Chain Rule & Bayes Rule

Chain rule:

\[ P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \]

Bayes rule:

\[ P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \]
Recap: Bayesian Learning

D is the measured data
Our goal is to estimate parameter $\theta$

• 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
Recap: MAP estimation for Binomial distribution

In the coin flip problem:

Likelihood is Binomial: \[ P(D \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \]

If the prior is Beta:

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

then the posterior is Beta distribution
Recap: Beta conjugate prior

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

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

As we get more samples, effect of prior is “washed out”
Application of Bayes Rule
AIDS test (Bayes rule)

### Data
- Approximately 0.1% are infected
- Test detects all infections
- Test reports positive for 1% healthy people

### Probability of having AIDS if test is positive

\[
P(a = 1|t = 1) = \frac{P(t = 1|a = 1)P(a = 1)}{P(t = 1)}
\]

\[
= \frac{P(t = 1|a = 1)P(a = 1)}{P(t = 1|a = 1)P(a = 1) + P(t = 1|a = 0)P(a = 0)}
\]

\[
= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091
\]

Only 9%!
Improving the diagnosis

Use a weaker follow-up test!

- Approximately 0.1% are infected
- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

\[
P(a = 0 | t_1 = 1, t_2 = 1) = \frac{P(t_1 = 1, t_2 = 1 | a = 0)P(a = 0)}{P(t_1 = 1, t_2 = 1 | a = 1)P(a = 1) + P(t_1 = 1, t_2 = 1 | a = 0)P(a = 0)} \\
= \frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357
\]

\[
P(a = 1 | t_1 = 1, t_2 = 1) = 0.643
\]

64%!
Why can’t we use Test 1 twice?

- Outcomes are not independent,
- But tests 1 and 2 conditionally independent (by assumption):

\[ p(t_1, t_2 | a) = p(t_1 | a) \cdot p(t_2 | a) \]
The Naïve Bayes Classifier
Data for spam filtering

- date
- time
- recipient path
- IP number
- sender
- encoding
- many more features

Delivered-To: alex.smola@gmail.com
Received: by 10.216.47.73 with SMTP id s51cs361171web;
  Tue, 3 Jan 2012 14:17:53 -0800 (PST)
Received: by 10.213.17.145 with SMTP id s17mr2519891eba.147.1325629071725;
  Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Return-Path: <alex+caf_=alex.smola@gmail.com>
Received: from mail-eY0-f175.google.com (mail-eY0-f175.google.com [209.85.215.175])
  by mx.google.com with ESMTPS id n4si29264232eef.57.2012.01.03.14.17.51
    (version=TLSv1/SSLv3 cipher=OTHER);
  Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Received-SPF: neutral (google.com: 209.85.215.175 is neither permitted nor denied by best
    guess record for domain of alex+caf_=alex.smola@gmail.com) client-ip:209.85.220.179;
  for <alex.smola@gmail.com>; Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Received: by 10.205.135.18 with SMTP id l1so15092746eaa.6
  for <alex.smola@gmail.com> (version=TLSv1/SSLv3 cipher=OTHER);
  Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Return-Path: <alex.smola@gmail.com>
Received: by eaal1 with SMTP id l1so15092746eaa.6
  for <alex.smola@gmail.com> (version=TLSv1/SSLv3 cipher=OTHER);
  Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Received: by 10.220.17.129 with HTTP; Tue, 3 Jan 2012 14:17:47 -0800 (PST)
Date: Tue, 3 Jan 2012 14:17:47 -0800 (PST)
X-Goog-Sender: <althoff.tim@googlemail.com>
Message-ID: <CAFJ3H4GDPW-5dZg9WMaABiAKyvidKjtPneModjiY6joG0-WC7osq@mail.gmail.com>
Subject: CS 281B. Advanced Topics in Learning and Decision Making
From: Tim Althoff, althoff@eecs.berkeley.edu
Naïve Bayes Assumption: Features $X_1$ and $X_2$ are conditionally independent given the class label $Y$:

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

More generally:

$$P(X_1...X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$
Task: Predict whether or not a picnic spot is enjoyable

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

<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>

Naïve Bayes assumption: \( P(X_1 \ldots X_d|Y) = \prod_{i=1}^{d} P(X_i|Y) \)

How many parameters to estimate?

(X is composed of \( d \) binary features, \( Y \) has \( K \) possible class labels)

\((2^d-1)K \) vs \((2-1)dK\)
Naïve Bayes Classifier

Given:

- Class prior $P(Y)$
- $d$ conditionally independent features $X_1, \ldots, X_d$ given the class label $Y$
- For each $X_i$ feature, we have the conditional likelihood $P(X_i | Y)$

Naïve Bayes Decision rule:

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

$$= \arg \max_y \prod_{i=1}^{d} P(x_i | y)P(y)$$
Naïve Bayes Algorithm for discrete features

Training data: \( \{(X^{(j)}, Y^{(j)})\}_{j=1}^{n} \)

\( X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)}) \)

\( n \) \( d \)-dimensional discrete features + \( K \) class labels

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

We need to estimate these probabilities!

Estimate them with MLE (Relative Frequencies)!
Naïve Bayes Algorithm for discrete features

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

We need to estimate these probabilities!

**Estimators**

For Class Prior

\[ \hat{P}(y) = \frac{\#j : Y(j) = y}{n} \]

For Likelihood

\[ \frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\#j : X_i(j) = x_i, Y(j) = y}{\#j : Y(j) = y} / n \]

**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: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

For example,

there is no $X_1 = \text{‘Earn’}$ when $Y = \text{‘SpamEmail’}$ in our dataset.

$\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a | Y = b) = 0$

$\Rightarrow P(X_1 = a, X_2 \ldots X_n | Y) = P(X_1 = a | Y) \prod_{i=2}^{d} P(X_i | Y) = 0$

Thus, no matter what the values $X_2, \ldots, X_d$ take:

$P(Y = b | X_1 = a, X_2, \ldots, X_d) = 0$

What now???
Naïve Bayes Alg — Discrete features

Training data:\n\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n} \quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})

Use your expert knowledge & apply prior distributions:
- Add m “virtual” examples
- Same as assuming conjugate priors

Assume priors:\n\[ Q(Y = b) \quad Q(X_i = a, Y = b) \]

MAP Estimate:\n\[ \hat{P}(X_i = a | Y = b) = \frac{\#j : X_i^{(j)} = a, Y^{(j)} = b + mQ(X_i = a, Y = b)}{\#j : Y^{(j)} = b + mQ(Y = b)} \]

called Laplace smoothing
Case Study: Text Classification
Is this spam?

BBM 406 on Piazza
To: Aykut Erdem
Instr Note] BBM 409 - Assignment 1

Instructor Aysun Koçak posted a new Note.

BBM 409 - Assignment 1

Hello,

Your first assignment is attached below. Dataset and movie information is given in assignment.zip.

pset1.pdf
assignment1.zip

Click here to view. Search or link to this question with @7.

Sign up for more classes at http://piazza.com/hacettepe.edu.tr.

Tell a colleague about Piazza. It's free, after all.

Thanks,
The Piazza Team

Contact us at team@piazza.com

You're receiving this email because aykut@cs.hacettepe.edu.tr is enrolled in BBM 406 at Hacettepe University. Sign in to manage your email preferences or un-enroll from this class.
Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.
What is the subject of this article?

MEDLINE Article

MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

slide by Dan Jurafsky
Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...

slide by Dan Jurafsky
Text Classification: definition

• Input:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

• Output: a predicted class $c \in C$
Hand-coded rules

• Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)

• Accuracy can be high
  - If rules carefully refined by expert

• But building and maintaining these rules is expensive
Text Classification and Naive Bayes

- Classify emails
  - \( Y = \{\text{Spam, NotSpam}\} \)

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

What are the features \( X \)?

The text!

Let \( X_i \) represent \( i^{th} \) word in the document
$X_i$ represents $i^{th}$ word in document

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
A problem: The support of $P(X|Y)$ is huge!

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

$$X_i \in \{1,...,50000\} \Rightarrow K(1000^{50000} - 1)$$

parameters to estimate without the NB assumption....

$$h_{MAP}(x) = \arg \max_{1 \leq k \leq K} P(Y = k) P(X_1 = x_1, \ldots, X_{1000} = x_{1000} | Y = k)$$
NB for Text Classification

\[ X_i \in \{1, \ldots, 50000\} \Rightarrow K(1000^{50000} - 1) \text{ parameters to estimate} \ldots \]

**NB assumption helps a lot!!!**

If \( P(X_i = x_i | Y=y) \) is the probability of observing word \( x_i \) at the \( i^{th} \) position in a document on topic \( y \)

\[ \Rightarrow 1000K(50000-1) \text{ parameters to estimate with NB assumption} \]

**NB assumption helps, but still lots of parameters to estimate.**

\[ h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{\text{LengthDoc}} P(X_i = x_i | y) \]
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
  The document is just a bag of words: i.i.d. words
- Sounds really silly, but often works very well!

\[\Rightarrow K(50000-1) \text{ parameters to estimate}\]

The probability of a document with words \(x_1, x_2, \ldots\)

\[
\prod_{i=1}^{\text{LengthDoc}} P(x_i | y) = \prod_{w=1}^{W} P(w | y)^{\text{count}_w}
\]
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.
The bag of words representation: using a subset of words

\[ \gamma(x) = c \]

- love
- sweet
- satirical
- great
- fun
- whimsical
- romantic
- laughing
- recommend
- several
- happy
- again
The bag of words representation

\[ \gamma( \text{great} ) = 2 \]
\[ \gamma( \text{love} ) = 2 \]
\[ \gamma( \text{recommend} ) = 1 \]
\[ \gamma( \text{laugh} ) = 1 \]
\[ \gamma( \text{happy} ) = 1 \]

...
Choosing a class:

\[
\hat{P}(c) = \frac{N_c}{N}
\]

\[
\hat{P}(w | c) = \frac{\text{count}(w,c) + 1}{\text{count}(c) + |V|}
\]

Priors:

\[
P(c) = \frac{3}{4}
\]

\[
P(j) = \frac{1}{4}
\]

Conditional Probabilities:

\[
P(\text{Chinese} | c) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}
\]

\[
P(\text{Tokyo} | c) = \frac{0+1}{8+6} = \frac{1}{14}
\]

\[
P(\text{Japan} | c) = \frac{0+1}{8+6} = \frac{1}{14}
\]

\[
P(\text{Chinese} | j) = \frac{1+1}{3+6} = \frac{2}{9}
\]

\[
P(\text{Tokyo} | j) = \frac{1+1}{3+6} = \frac{2}{9}
\]

\[
P(\text{Japan} | j) = \frac{1+1}{3+6} = \frac{2}{9}
\]

Choosing a class:

\[
P(\text{cld5}) \propto \frac{3}{4} \times \left(\frac{3}{7}\right)^3 \times \frac{1}{14} \times \frac{1}{14} \approx 0.0003
\]

\[
P(\text{jld5}) \propto \frac{1}{4} \times \left(\frac{2}{9}\right)^3 \times \frac{2}{9} \times \frac{2}{9} \approx 0.0001
\]
What if features are continuous?

e.g., 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

\[ h_{NB}(x) = \arg \max_y P(y) \prod_i P(X_i = x_i | y) \]
\[ \approx \arg \max_k \hat{P}(Y = k) \prod_i \mathcal{N}(\hat{\mu}_{ik}, \hat{\sigma}_{ik}) \]

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

\[ \hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{j=1}^{N} (x_j - \hat{\mu})^2 \]
Estimating parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

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

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

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

\[ \hat{\sigma}^2_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X^j_i - \hat{\mu}_{ik})^2 \delta(Y^j = y_k) \]
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  misc.forsale
comp.os.ms-windows.misc  rec.autos
comp.sys.ibm.pc.hardware  rec.motorcycles
comp.sys.mac.hardware  rec.sport.baseball
comp.windows.x  rec.sport.hockey

alt.atheism  sci.space
soc.religion.christian  sci.crypt
talk.religion.misc  sci.electronics
talk.politics.mideast  sci.med
talk.politics.misc
talk.politics.guns

Naïve Bayes: 89% accuracy
Case Study:
Classifying Mental States
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.]
Brain scans can track activation with precision and sensitivity.
Learned Naïve Bayes Models – Means for $P(\text{BrainActivity} \mid \text{WordCategory})$

Pairwise classification accuracy: 78-99%, 12 participants

[Mitchell et al.]
What you should know…

Naïve Bayes classifier
- What’s the assumption
- Why we use it
- How do we learn it
- Why is Bayesian (MAP) estimation important

Text classification
- Bag of words model

Gaussian NB
- Features are still conditionally independent
- Each feature has a Gaussian distribution given class