Naïve Bayes, Gaussian Distributions, Practical Applications

Required reading:
• Mitchell draft chapter, sections 1 and 2.
  (available on class website)

Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

Ground Hog’s Day, 2009
Overview

Recently:
• learn $P(Y|X)$ instead of deterministic $f:X \rightarrow Y$
• Bayes rule
• MLE and MAP estimates for parameters of $P$
• Conditional independence
• classification with Naïve Bayes

Today:
• Text classification with Naïve bayes
• Gaussian distributions for continuous $X$
• Gaussian Naïve Bayes classifier
• Image classification with Naïve bayes
Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?
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
Baseline: Bag of Words Approach

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.

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>0</td>
</tr>
<tr>
<td>about</td>
<td>2</td>
</tr>
<tr>
<td>all</td>
<td>2</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
</tr>
<tr>
<td>apple</td>
<td>0</td>
</tr>
<tr>
<td>anxious</td>
<td>0</td>
</tr>
<tr>
<td>gas</td>
<td>1</td>
</tr>
<tr>
<td>oil</td>
<td>1</td>
</tr>
<tr>
<td>Zaire</td>
<td>0</td>
</tr>
</tbody>
</table>
Naïve Bayes in a Nutshell

Bayes rule:

\[ P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k)P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1 \ldots X_n | Y = y_j)} \]

Assuming conditional independence among \( X_i \)'s:

\[ P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)} \]

So, classification rule for \( X^{\text{new}} = < X_1, \ldots, X_n > \) is:

\[ Y^{\text{new}} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{\text{new}} | Y = y_k) \]

\( X_i \in \{1, 0\} \rightarrow \Theta_i = P(X_i = 1) \quad (1 - \Theta_i) = P(X_i = 0) \)

\( X_i = \text{count of word } i \in \{0, 1, \ldots \} \)
Learning to Classify Text

Target concept *Interesting?* : *Document* → {+, −}

1. Represent each document by vector of words
   - one attribute per word position in document

2. Learning: Use training examples to estimate
   - \( P(+) \)
   - \( P(−) \)
   - \( P(doc|+) \)
   - \( P(doc|−) \)

Naive Bayes conditional independence assumption

\[
P(doc|v_j) = \prod_{i=1}^{\text{length}(doc)} P(a_i = w_k|v_j)
\]

where \( P(a_i = w_k|v_j) \) is probability that word in position \( i \) is \( w_k \), given \( v_j \)

one more assumption:

\[
P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m
\]
Twenty Newsgroups

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

Naive Bayes: 89% classification accuracy
LEARN_NAIVE_BAYES_TEXT(Examples, V)

1. collect all words and other tokens that occur in Examples

- Vocabulary ← all distinct words and other tokens in Examples

2. calculate the required $P(v_j)$ and $P(w_k|v_j)$ probability terms

- For each target value $v_j$ in V do

  - $docs_j$ ← subset of Examples for which the target value is $v_j$
  - $P(v_j) ← \frac{|docs_j|}{|Examples|}$
  - $Text_j$ ← a single document created by concatenating all members of $docs_j$
  - $n ←$ total number of words in $Text_j$ (counting duplicate words multiple times)
  - for each word $w_k$ in Vocabulary
      - $n_k ←$ number of times word $w_k$ occurs in $Text_j$
      - $P(w_k|v_j) ← \frac{n_k+1}{n+|Vocabulary|}$

For code and data, see www.cs.cmu.edu/~tom/mlbook.html click on "Software and Data"
classify_naive_bayes_text(Doc)

- \textit{positions} \leftarrow \text{all word positions in } Doc \text{ that contain tokens found in } Vocabulary

- Return \( v_{NB} \), where

\[
v_{NB} = \arg \max_{v_j \in V} \underbrace{P(v_j)}_{v_j \in \text{positions}} \prod_{i \in \text{positions}} P(a_i|v_j)
\]

\[
v_{NB} = \arg \max_{v_j \in V} \log (P(v_j)) + \sum_{i} \log P(a_i|v_j) \geq \log P(c|v_j)
\]

\[
P(v = v_j | Doc) = \frac{P(v_j) \prod_{i} P(a_i | v_j)}{P(Doc)}
\]
Learning Curve for 20 Newsgroups

Accuracy vs. Training set size (1/3 withheld for test)
What if we have continuous $X_i$?

Eg., image classification: $X_i$ is real-valued $i^{th}$ pixel
What if we have continuous $X_i$?

Eg., image classification: $X_i$ is real-valued $i^{th}$ pixel

Naïve Bayes requires $P(X_i \mid Y_k)$, but $X_i$ is real (continuous)

$$P(Y \mid x) = \frac{P(Y) P(x \mid Y)}{P(x)} \quad \text{(c.i.)} \quad P(Y) \prod_{i} P(X_i \mid Y_k) \quad \text{(Gaussian)}$$

Common approach: assume $P(X_i \mid Y_k)$ follows a normal (Gaussian) distribution
Gaussian Distribution
(also known as “Normal” distribution)

\[ p(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2} \]

The probability that \( X \) will fall into the interval \((a, b)\) is given by

\[ P(a < X < b) = \int_{a}^{b} p(x) \, dx \]

- Expected, or mean value of \( X \), \( E[X] \), is
  \[ E[X] = \mu \]

- Variance of \( X \) is
  \[ Var(X) = \sigma^2 \]

- Standard deviation of \( X \), \( \sigma_X \), is
  \[ \sigma_X = \sigma \]
What if we have continuous $X_i$?

Eg., image classification: $X_i$ is $i^{th}$ pixel

Gaussian Naïve Bayes (GNB): assume

$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi \sigma_{ik}^2}} e^{-\frac{1}{2} \left( \frac{x - \mu_{ik}}{\sigma_{ik}} \right)^2}$$

Sometimes assume variance

- is independent of $Y$ (i.e., $\sigma_j$),
- or independent of $X_i$ (i.e., $\sigma_k$),
- or both (i.e., $\sigma$)
Gaussian Naïve Bayes Algorithm – continuous $X_i$
(but still discrete $Y$)

- Train Naïve Bayes (examples)
  for each value $y_k$
  estimate* $\pi_k \equiv P(Y = y_k)$
  for each attribute $X_i$ estimate
  class conditional mean $\mu_{ik}$, variance $\sigma_{ik}$

- Classify ($X^{new}$)

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

  $Y^{new} \leftarrow \arg \max_{y_k} \pi_k \prod_i \text{Normal}(X_i^{new}, \mu_{ik}, \sigma_{ik})$

* probabilities must sum to 1, so need estimate only n-1 parameters...
Estimating Parameters: $Y$ discrete, $X_i$ continuous

Maximum likelihood estimates:

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

$\delta(z) = 1$ if $z$ true, else 0

$$\hat{\sigma}^2_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X^j_i - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$
How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, X=<X1, … Xn>?
What is form of decision surface for Gaussian Naïve Bayes classifier?

eg., if distributions are spherical, attributes have same variance

\( \sigma_{ik} = \sigma \)
What is form of decision surface for Naïve Bayes classifier?
GNB Example: Classify a person’s cognitive activity, based on brain image

- reading a sentence or viewing a picture?
- reading the word “Hammer” or “Apartment”
- viewing a vertical or horizontal line?
- answering the question, or getting confused?
Stimuli for our study:

60 distinct exemplars, presented 6 times each
fMRI voxel means for “bottle”: means defining $P(X_i | Y=\text{“bottle”})$

Mean fMRI activation over all stimuli:

“bottle” minus mean activation:
Training Classifiers over fMRI sequences

• Learn the classifier function
  \[ \text{Mean}(\text{fMRI}(t+4), \ldots, \text{fMRI}(t+7)) \rightarrow \text{WordCategory} \]
  - Leave one out cross validation

• Preprocessing:
  - Adjust for head motion
  - Convert each image \( x \) to standard normal image
    \[
    x(i) \leftarrow \frac{x(i) - \mu_x}{\sigma_x}
    \]

• Learning algorithms tried:
  - kNN (spatial correlation)
  - SVM
  - SVDM
  - Gaussian Naïve Bayes
  - Regularized Logistic regression

• Feature selection methods tried:
  - Logistic regression weights, voxel stability, activity relative to fixation,...
Classification task: is person viewing a “tool” or “building”?

![Bar chart showing classification accuracy for participants p4 to p1. Classification accuracy is statistically significant with p<0.05.](image)
Where in the brain is activity that distinguishes tools vs. buildings?

Accuracy of a radius one classifier centered at each voxel:
voxel clusters: searchlights

Accuracies of cubical 27-voxel classifiers centered at each significant voxel [0.7-0.8]
Are classifiers detecting neural representations of meaning or perceptual features?

ML: Can we train on word stimuli, then decode picture stimuli?

**YES**: We can train classifiers when presenting English words, then decode category of picture stimuli, or Portuguese words.

Therefore, the learned neural activation patterns must capture how the brain represents the meaning of input stimulus.
Are representations similar across different people?

ML: Can we train classifier on data from a collection of people, then decode stimuli for a new person?

**YES**: We can train on one group of people, and classify fMRI images of new person

Therefore, seek a theory of neural representations common to all of us (and of how we vary)
What you should know:

• Training and using classifiers based on Bayes rule

• Conditional independence
  – What it is
  – Why it’s important

• Naïve Bayes
  – What it is
  – Why we use it so much
  – Training using MLE, MAP estimates
  – Discrete variables (Bernoulli) and continuous (Gaussian)
Questions to think about:

- Can you use Naïve Bayes for a combination of discrete and real-valued $X_i$?

- How can we easily model just 2 of n attributes as dependent?

- What does the decision surface of a Naïve Bayes classifier look like?

- How would you select a subset of $X_i$’s?