Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 28, 2015

Today:
• Naïve Bayes
  • discrete-valued $X_i$’s
  • Document classification
• Gaussian Naïve Bayes
  • real-valued $X_i$’s
  • Brain image classification

Readings:

Required:
• Mitchell: “Naïve Bayes and Logistic Regression”
  (available on class website)

Optional
• Bishop 1.2.4
• Bishop 4.2
Recently:
• Bayes classifiers to learn $P(Y|X)$
• MLE and MAP estimates for parameters of $P$
• Conditional independence
• Naïve Bayes \(\rightarrow\) make Bayesian learning practical

Next:
• Text classification
• Naïve Bayes and continuous variables $X_i$:
  • Gaussian Naïve Bayes classifier
• Learn $P(Y|X)$ directly
  • Logistic regression, Regularization, Gradient ascent
• Naïve Bayes or Logistic Regression?
  • Generative vs. Discriminative classifiers
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 \( X_{\text{new}} = \langle X_1, \ldots, X_n \rangle \) is:

\[ Y_{\text{new}} \leftarrow \arg \max_{y_k} P(Y = y_k)\prod_i P(X_i^{\text{new}}|Y = y_k) \]
Example: Live in Sq Hill? $P(S|G,D,B)$ $n = 18 + 33 + 22 + 29$

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive or Carpool to CMU
- $B=1$ iff Birthday is before July 1

\[P(S=1) = \frac{5 + 7 + 10 + 4}{26} = \frac{26}{102}\]
\[P(S=0) = \frac{76}{102}\]

\[P(D=1 | S=1) = \frac{3}{26}\]
\[P(D=0 | S=1) = \frac{23}{26}\]
\[P(D=1 | S=0) = \frac{1}{76}\]
\[P(D=0 | S=0) = \frac{76}{76}\]

\[P(G=1 | S=1) = \frac{5 + 8 + 4 + 9}{26} = \frac{26}{26}\]
\[P(G=1 | S=0) = \frac{7}{102}\]
\[P(G=0 | S=1) = \frac{0}{26}\]
\[P(G=0 | S=0) = \frac{26}{76}\]

\[P(B=1 | S=1) = \frac{1 + 2 + 5 + 2}{26} = \frac{10}{26}\]
\[P(B=0 | S=1) = \frac{16}{26}\]
\[P(B=1 | S=0) = \frac{5 + 7 + 8 + 6}{76} = \frac{26}{76}\]
\[P(B=0 | S=0) = \frac{50}{76}\]

Tom: $D=1$, $G=0$, $B=0$

\[P(S=1|D=1,G=0,B=0) = \frac{P(S=1) P(D=1|S=1) P(G=0|S=1) P(B=0|S=1)}{P(S=1) P(D=1|S=1) P(G=0|S=1) P(B=0|S=1) + P(S=0) P(D=1|S=0) P(G=0|S=0) P(B=0|S=0)}\]
Another way to view Naïve Bayes (Boolean Y): 
Decision rule: is this quantity greater or less than 1?

\[
\frac{P(Y = 1|X_1 \ldots X_n)}{P(Y = 0|X_1 \ldots X_n)} = \frac{P(Y = 1) \prod_i P(X_i|Y = 1)}{P(Y = 0) \prod_i P(X_i|Y = 0)}
\]
Naïve Bayes: classifying text documents

• Classify which emails are spam?
• Classify which emails promise an attachment?

How shall we represent text documents for Naïve Bayes?

I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.

I would like to thank Frank Pfenning, who has served ably in this role for the past two years.

________________________
Randal E. Bryant
Dean and University Professor
Learning to classify documents: $P(Y|X)$

- $Y$ discrete valued.
  - e.g., Spam or not
- $X = <X_1, X_2, \ldots X_n> = $ document

- $X_i$ is a random variable describing...
Learning to classify documents: $P(Y|X)$

- $Y$ discrete valued.
  - e.g., Spam or not
- $X = <X_1, X_2, \ldots X_n> = \text{document}$

- $X_i$ is a random variable describing…

Answer 1: $X_i$ is boolean, 1 if word $i$ is in document, else 0
  e.g., $X_{\text{pleased}} = 1$

Issues?
Learning to classify documents: $P(Y|X)$

- $Y$ discrete valued.
  - e.g., Spam or not
- $X = <X_1, X_2, \ldots, X_n> = $ document

- $X_i$ is a random variable describing…
Answer 2:
- $X_i$ represents the $i^{th}$ word position in document
- $X_1 = “I”, \ X_2 = “am”, \ X_3 = “pleased”$
- and, let’s assume the $X_i$ are iid (indep, identically distributed)

\[ P(X_i|Y) = P(X_j|Y) \quad (\forall i, j) \]
Learning to classify document: $P(Y|X)$
the “Bag of Words” model

- $Y$ discrete valued. e.g., Spam or not
- $X = <X_1, X_2, \ldots, X_n> = \text{document}$

- $X_i$ are iid random variables. Each represents the word at its position $i$ in the document
- Generating a document according to this distribution = rolling a 50,000 sided die, once for each word position in the document

- The observed counts for each word follow a ??? distribution
Multinomial Distribution

Eg. 2 Dice roll problem (6 outcomes instead of 2)

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

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

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

\[
P(\theta \mid D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)
\]

\[
\hat{\theta}_i^{\text{MAP}} = \hat{P}(X = i) = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{k} (\alpha_j + \beta_j - 1)}
\]
Multinomial Bag of Words

<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>
MAP estimates for bag of words

Map estimate for multinomial

\[
\hat{\theta}_i^{MAP} = \hat{P}(X = i) = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{k}(\alpha_j + \beta_j - 1)}
\]

\[
\hat{\theta}_{\text{aardvark}}^{MAP} = P(X = \text{aardvark}) = \frac{\# \text{ observed 'aardvark'} + \# \text{ hallucinated 'aardvark'}}{\# \text{ observed words} + \# \text{ hallucinated words}}
\]

What β’s should we choose?
**Naïve Bayes Algorithm – discrete $X_i$**

- **Train Naïve Bayes (examples)**
  - for each value $y_k$
    - estimate $\pi_k \equiv P(Y = y_k)$
  - for each value $x_{ij}$ of each attribute $X_i$
    - estimate $\theta_{ijk} \equiv P(X_i = x_{ij}|Y = y_k)$

- **Classify ($X^{new}$)**
  
  $$
  Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_{i}^{new}|Y = y_k) \\
  Y^{new} \leftarrow \arg \max_{y_k} \pi_k \prod_i \theta_{ijk}
  $$

* Additional assumption: word probabilities are position independent

$$
\theta_{ijk} = \theta_{mjk} \text{ for } i \neq m
$$
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
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 $i^{\text{th}}$ pixel
What if we have continuous $X_i$?

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

Naïve Bayes requires $P(X_i \mid Y=y_k)$, but $X_i$ is real (continuous)

$$P(Y = y_k \mid X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i \mid Y = y_j)}$$

Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution
What if we have continuous $X_i$?

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

Naïve Bayes requires $P(X_i \mid Y=y_k)$, but $X_i$ is real (continuous)

$$P(Y = y_k \mid X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i \mid Y = y_j)}$$

Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution
Gaussian Distribution (also called “Normal”)

$p(x)$ is a **probability density function**, whose integral (not sum) is 1

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

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

\[ \int_{a}^{b} p(x) \, dx \]

- 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 $X_i$?

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., $\sigma_i$),
- or independent of $X_i$ (i.e., $\sigma_k$)
- or both (i.e., $\sigma$)
Train Naïve Bayes (examples) for each value \( y_k \)

estimate* \( \pi_k \equiv P(Y = y_k) \)

for each attribute \( X_i \) estimate \( P(X_i|Y = y_k) \)

• class conditional mean \( \mu_{ik} \), variance \( \sigma_{ik} \)

Classify \( (X^{new}) \)

\[ Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_{i} P(X^{new}_i|Y = y_k) \]

\[ Y^{new} \leftarrow \arg \max_{y_k} \pi_k \prod_{i} \mathcal{N}(X^{new}_i; \mu_{ik}, \sigma_{ik}) \]

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

Maximum likelihood estimates:

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

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

\(\delta() = 1\) if \((Y^j = y_k)\), else 0
How many parameters must we estimate for Gaussian Naïve Bayes if $Y$ has $k$ possible values, $X=<X_1, \ldots, X_n>$?

\[ p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2} \]
GNB Example: Classify a person’s cognitive state, based on brain image

- reading a sentence or viewing a picture?
- reading the word describing a “Tool” or “Building”?
- answering the question, or getting confused?
Mean activations over all training examples for Y=“bottle”

Y is the mental state (reading “house” or “bottle”)

Xᵢ are the voxel activities,

this is a plot of the μ’s defining $P(Xᵢ \mid Y=\text{“bottle”})$
Classification task: is person viewing a “tool” or “building”?

Participants

Classification accuracy

statistically significant

p<0.05
Where is information encoded in the brain?

Accuracies of cubical 27-voxel classifiers centered at each significant voxel [0.7-0.8]
Naïve Bayes: What you should know

• Designing classifiers based on Bayes rule

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

• Naïve Bayes assumption and its consequences
  – Which (and how many) parameters must be estimated under different generative models (different forms for \( P(X|Y) \) )
    • and why this matters

• How to train Naïve Bayes classifiers
  – MLE and MAP estimates
  – with discrete and/or continuous inputs \( X_i \)
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?
Simple Picture for GNB for $P(Y|X_1)$