Bayes optimal classifier
Naïve Bayes

Machine Learning – CSE446
Carlos Guestrin
University of Washington
May 24, 2013

Classification

- **Learn:** h:X ↦ Y
  - X – features
  - Y – target classes

- Suppose you know P(Y|X) exactly, how should you classify?
  - Bayes optimal classifier:
Bayes Rule

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

Which is shorthand for:

\[ (\forall i, j) P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{P(X = x_j)} \]

How hard is it to learn the optimal classifier?

- Data =
  - How do we represent these? How many parameters?
    - Prior, \( P(Y) \):
      - Suppose \( Y \) is composed of \( k \) classes
    - Likelihood, \( P(X|Y) \):
      - Suppose \( X \) is composed of \( d \) binary features

- Complex model ! High variance with limited data!!!
Conditional Independence

- X is **conditionally independent** of Y given Z, if the probability distribution governing 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., \(P(\text{Thunder}|\text{Rain, Lightning}) = P(\text{Thunder}|\text{Lightning})\)

- Equivalent to:

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

What if features are independent?

- Predict Thunder
- From two **conditionally Independent** features
  - Lightening
  - Rain
The 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:
    \[ P(X_1...X_d|Y) = \prod_{i} P(X_i|Y) \]

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

The Naïve Bayes Classifier

- Given:
  - 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:
  \[ y^* = h_{NB}(x) = \arg \max_y P(y)P(x_1,\ldots,x_d | y) = \arg \max_y P(y) \prod_i P(x_i|y) \]

- If assumption holds, NB is optimal classifier!
MLE for the parameters of NB

- Given dataset
  - Count(A=a,B=b) == number of examples where A=a and B=b

- MLE for NB, simply:
  - Prior: \( P(Y=y) = \)
  - Likelihood: \( P(X_i=x|Y=y) = \)

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:
  \[
  P(X_1 \ldots X_d | Y) \neq \prod_i P(X_i | Y)
  \]

- Actual probabilities \( P(Y|X) \) often biased towards 0 or 1
- 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
Subtleties of NB classifier 2 – Insufficient training data

- What if you never see a training instance where \( X_1 = a \) when \( Y = b \)?
  - e.g., \( Y = \{ \text{SpamEmail} \}, X_1 = \{ \text{Enlargement} \} \)
  - \( P(X_1 = a \mid Y = b) = 0 \)
- Thus, no matter what the values \( X_2, \ldots, X_d \) take:
  - \( P(Y = b \mid X_1 = a, X_2, \ldots, X_d) = 0 \)

- “Solution”: smoothing
  - Add “fake” counts, usually uniformly distributed
  - Equivalent to Bayesian Learning

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!
Features $X$ are entire 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 Alexei 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-there-in-Toronto decided

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 $x_i$ 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!

\[
P(y) \prod_{i=1}^{Length\,Doc} P(x_i | y)
\]

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.
Bag of Words Approach

NB with Bag of Words for text classification

- Learning phase:
  - Prior $P(Y)$
    - Count how many documents you have from each topic (+ prior)
  - $P(X_i|Y)$
    - For each topic, count how many times you saw word in documents of this topic (+ prior)

- Test phase:
  - For each document
    - Use naïve Bayes decision rule

$$h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$
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
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- alt.atheism
- soc.religion.christian
- talk.religion.misc
- talk.politics.mideast
- talk.politics.misc
- talk.politics.guns
- sci.space
- sci.crypt
- sci.electronics
- sci.med

Naive Bayes: 89% classification accuracy

Learning curve for Twenty News Groups

Accuracy vs. Training set size (1/3 withheld for test)
Bayesian Networks – Representation

Machine Learning – CSE446
Carlos Guestrin
University of Washington
May 24, 2013

©Carlos Guestrin 2005-2013

Handwriting recognition

Character recognition, e.g., kernel SVMs
Webpage classification

Company home page vs
Personal home page vs
University home page vs
...

Handwriting recognition 2
Today – Bayesian networks

- One of the most exciting advancements in statistical AI in the last decades
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies
Causal structure

- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?

Possible queries

- Inference
- Most probable explanation
- Active data collection
Car starts BN

- 18 binary attributes
- Inference
  - $P(\text{BatteryAge}|\text{Starts}=f)$

- $2^{16}$ terms, why so fast?
- Not impressed?
  - HailFinder BN – more than $3^{54} = 58149737003040059690390169$ terms

Factored joint distribution - Preview

- Flu
- Allergy
- Sinus
- Headache
- Nose
What about probabilities?
Conditional probability tables (CPTs)

Number of parameters
Key: Independence assumptions

Knowing sinus separates the variables from each other

(Marginal) Independence

- Flu and Allergy are (marginally) independent
Marginally independent random variables

- **Sets** of variables $X$, $Y$
- $X$ is independent of $Y$ if
  - $P \perp (X=x \perp Y=y), \forall x \in \text{Val}(X), y \in \text{Val}(Y)$

- Shorthand:
  - Marginal independence: $P \perp (X \perp Y)$

- **Proposition**: $P$ satisfies $(X \perp Y)$ if and only if
  - $P(X,Y) = P(X) \cdot P(Y)$

Conditional independence

- Flu and Headache are not (marginally) independent

- Flu and Headache are independent given Sinus infection

- More Generally:
Conditionally independent random variables

- **Sets** of variables $X$, $Y$, $Z$
- $X$ is independent of $Y$ given $Z$ if
  \[ P(\{X=x, Y=y|Z=z\}) \forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z) \]

- **Shorthand:**
  - Conditional independence: $P \vdash (X \perp Y | Z)$
  - For $P \vdash (X \perp Y | \emptyset)$, write $P \vdash (X \perp Y)$

- **Proposition:** $P$ statisfies $(X \perp Y | Z)$ if and only if
  \[ P(X,Y|Z) = P(X|Z) P(Y|Z) \]

---

The independence assumption

**Local Markov Assumption:**
A variable $X$ is independent of its non-descendants given its parents
Explaining away

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents

Flu — Allergy — Sinus — Headache — Nose

Naïve Bayes revisited

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents
Joint distribution

Why can we decompose? Markov Assumption!

The chain rule of probabilities

- \( P(A,B) = P(A)P(B|A) \)

- More generally:
  \[
  P(X_1,\ldots,X_n) = P(X_1) \ P(X_2|X_1) \ \ldots \ \ P(X_n|X_1,\ldots,X_{n-1})
  \]
Chain rule & Joint distribution

Local Markov Assumption:
A variable $X$ is independent of its non-descendants given its parents

The Representation Theorem – Joint Distribution to BN

BN:

Encodes independence assumptions

If conditional independencies in BN are subset of conditional independencies in $P$

Obtain

Joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_{X_i})$$
Two (trivial) special cases

- Edgeless graph
- Fully-connected graph