Classification

Learn: \( f: X \mapsto Y \) → hired/not hired
- X - features
- Y - target classes

Suppose you know \( P(Y|X) \) exactly, how should you classify?
- Bayes optimal classifier:
  \[
  \hat{y} = \arg \max_y P(Y=y|X=x)
  \]
  In logistic regression, model for \( P(Y|X=x) = \frac{1}{1+e^{-wx}} \)
  Model \( P(Y|X) \) directly... “discriminative model”
Recall: Bayes rule

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

Which is shorthand for:

\[ (\forall i, j) \quad P(Y = i \mid X = j) = \frac{P(X = j \mid Y = i)P(Y = i)}{P(X = j)} \]

How hard is it to learn the optimal classifier?

- Data =

<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>Cold</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- How do we represent these? How many parameters?
  - Prior, \( P(Y) \):
    - Suppose \( Y \) is composed of \( k \) classes
      \[ k-1 \]
  - Likelihood, \( P(X \mid Y) \):
    - Suppose \( X \) is composed of \( d \) binary features
      \[ k(2^d - 1) \] a lot of params! need a lot of data
      \[ \sum_{y=1}^{k} P(Y=y) = 1 \]

- 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.

\[(\forall i, j, k) \quad P(X = i \mid Y = j, Z = k) = P(X = i \mid Z = k)\]

e.g.,

\[P(\text{Thunder} \mid \text{Rain, Lightening}) = P(\text{Thunder} \mid \text{Lightening})\]

\[\text{R LT? No! But R LT!} \]

Equivalent to:

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

What if features are independent?

• Predict Lightening
• From two conditionally independent features
  - Thunder
  - Rain

Estimate: \(P(X \mid Y) = P(T, R \mid L)\)

But \(T \perp R \mid L\)

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

\[= \frac{2 \cdot (2^2 - 1)}{2}\]

4 params!
The Naïve Bayes assumption

- Naïve Bayes assumption:
  - Features are independent given class:
    \[
    \]
  - More generally:
    \[
    P(X[1], \ldots, X[d] \mid Y) = \prod_{j=1}^{d} P(X[j] \mid Y)
    \]

- How many parameters now?
  - Suppose X is composed of d binary features
    \[\text{No assump.: } k(2^d - 1) \quad \text{Naïve Bayes: } k \cdot d\]
    \[\text{nice reduction! (might be too aggressive)}\]

The Naïve Bayes classifier

- Given:
  - Prior P(Y)
  - d conditionally independent features X[j] given the class Y
  - For each X[j], we have likelihood P(X[j]|Y)

- Decision rule:
  \[
  \hat{y} = f_{NB}(x) = \arg \max_{y} P(y)P(x[1], \ldots, x[d] \mid y)
  = \arg \max_{y} P(y)\prod_{j} P(x[j] \mid y)
  \]

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

- Given dataset
  - \( \text{Count}(A=a,B=b) == \# \text{ examples where } A=a \text{ and } B=b \)

- MLE for NB, simply:
  - Prior: \( P(Y=y) = \frac{\text{Count}(Y=y)}{\text{Count}(Y=y)} \)
    e.g. \( P(Y=\text{'hired'}) \)
  - Likelihood: \( P(X[j]=x[j] \mid Y=y) = \frac{\text{Count}(X=x,Y=y)}{\text{Count}(Y=y)} \)

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:
  \[ P(X[1], \ldots, X[d] \mid Y) \neq \prod_j P(X[j] \mid Y) \]

- Actual probabilities \( P(Y \mid X) \) often biased towards 0 or 1

- Nonetheless, NB is one of the 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., $Y=\{\text{SpamEmail}\}$, $X[1]=\{\text{Viagra}\}$
  - $P(X[1]=a \mid Y=b) = 0$

- 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! (very naively)
Features X are entire document – X[j] for jth word in article

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year’s biggest and worst (opin)
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-khuzai-in Toronto decided

NB for text classification

- P(X|Y) is huge!!!
  - Article at least 1000 words, X={X[1],..., X[1000]}
  - X[j] represents jth word in document
    - i.e., the domain of X[j] is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.

- NB assumption helps a lot!!!
  - P(X[j]=x[j]|Y=y) is the probability of observing word x[j] in a document on topic y

\[
f_{NB}(x) = \arg \max_y P(y) \prod_{j=1}^{\text{LengthDoc}} P(x[j] | y)
\]

©2017 Emily Fox
Bag of words model

- Typical additional assumption: Position in document doesn’t matter

\[ P(X[j] = x[j] | Y=y) = P(X[k] = x[j] | Y=y) \]

- “Bag of words” model – order of words on the page ignored
- Sounds really silly, but often works very well!

\[
\begin{align*}
P(y) \prod_{j=1}^{\text{LengthDoc}} P(x[j] | y)
\end{align*}
\]

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.
Bag-of-words representation

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Eugene F. Weber, Eunji M. R. Park, Brian Litt

Abstract

Patients with epilepsy can experience shift in clinical and subclinical seizure patterns. We developed a Bayesian framework for modeling the complex dynamics and changing correlations of EEG data that captures the interdependence between different time-series. The model is trained using a Markov chain Monte Carlo algorithm. The model is applied to a dataset of EEG recordings from patients with epilepsy. The model outperforms existing approaches in terms of accuracy and interpretability.

1. Introduction

Despite recent advances in research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disorder and the paucity of quantitative tools that are feasible...
NB with bag of words for text classification

- **Learning phase:**
  - Prior $P(Y)$
    - Count how many documents you have from each topic (+ prior)
  - $P(X[j]|Y)$
    - 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
      \[
      f_{NB}(x) = \arg \max_y P(y) \prod_{j=1}^{\text{LengthDoc}} P(x[j] | y)
      \]

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
- alt.atheism
- soc.religion.christian
- talk.religion.misc
- talk.politics.mideast
- talk.politics.misc
- talk.politics.guns
- misc.forsale
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- 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

CSE 446: Machine Learning
Emily Fox
University of Washington
March 3, 2017
Learning from structured data

TrueSkill: A Bayesian Skill Rating System

Herbrich et al., 2007
ICU Monitoring

Beinlich et al., 1989

Aleks, Russell, et al., 2008

Digging in:
Learning with and without context/structure
Without context: Handwriting recognition

Character recognition, e.g., kernel SVMs

Without context: Webpage classification

Company website

University website

Personal website

...
With context: Handwriting recognition

With context: Webpage classification
Modeling structured relationships via Bayesian networks

Today – Bayesian networks

- Provided a huge advancement in AI/ML
- Generalizes naïve Bayes and logistic regression
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies
Bayesian network representation

Compact representation of a probability distribution.

Directed Acyclic Graph

Vertices: Random Variables
Edges: Conditional dependencies
“probabilistic relationships”

Bayesian network probability factorization

One CPT (conditional probability table) for each variable

\[ P(\text{variable} \mid \text{parents of variable}) \]

implies the factorization:

\[ P(X) = \prod_{i=1}^{\left|X\right|} P(X_i \mid \text{parents}(X_i)) \]

\[ P(A,B,C,D) = P(A) \cdot P(B) \cdot P(C \mid A,B) \cdot P(D \mid C) \]
What a Bayesian network represents (in detail) and what does it buy you?

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

CarStarts? Bayesian network

- 18 binary attributes
- Inference
  - \( P(\text{BatteryAge}|\text{Starts}=\text{f}) \)
- \( 2^{16} \) terms, why so fast?
- Not impressed?
  - HailFinder BN – more than \( 3^{34} = 58149737003040059690390169 \) terms
Factored joint distribution – A preview

What are these probabilities? Conditional probability tables (CPTs)
Number of parameters

Flu → Sinus → Headache
Sinus → Allergy
Nose

Factorization speeds up inference

Exploit distributivity:

\[ P(F = z_F | N = t) \propto \sum_{x_A, x_S, x_H} P(F = x_F, A = x_A, S = x_S, H = x_H, N = t) \]

\[ = \sum_{x_A, x_S, x_H} P(F = x_F) P(A = x_A) P(S = x_S | F = x_F, A = x_A) P(H = x_H | S = x_S) P(N = t | S = x_S) \]

\[ = P(F = z_F) \sum_{x_A} P(A = x_A) \sum_{x_S} P(S = x_S | F = z_F, A = x_A) P(N = t | S = x_S) \sum_{x_H} P(H = x_H | S = x_S) \]
Key: Independence assumptions

Knowing sinus separates variables from each other

Marginal and conditional independence
(Marginal) Independence

- Flu and Allergy are (marginally) independent

<table>
<thead>
<tr>
<th>Flu</th>
<th>Allergy</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
</tr>
</tbody>
</table>

Conditional independence

- Flu and Headache are not (marginally) ind.

- Flu and Headache are independent given Sinus infection

- More generally:
Conditional independence statements encoded by Bayesian networks

What is a Bayes net assuming?

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

- $E \perp A \mid B, C$
- $E \perp D \mid B, C$
- $F \perp B \mid E$

Allows you to read off some simple conditional independence relationships
Conditional independence in Bayes nets

- Consider 4 different junction configurations

- Conditional versus unconditional independence:

Explaining away example

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

• Consider 4 different junction configurations

• Bayes ball algorithm:

Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H

V structure.
C not observed. Ball bounces away.
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H

V structure.
C observed. Ball can pass through
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H

Ball gets stuck here
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H