

Statistical Machine Learning Overview

Lecture 3: Slides from Tom Mitchell

Doug Aberdeen, October 14, 2004

National ICT Australia

News

- Canberra venue on the 25th and 28th will be in RSISE Rm 207
- First assignment is online
- *Mandatory* EM reading is online

MAP and ML

- MAP = maximum *a posteriori* (maximally probable)
 - $h_{map} = \arg \max_h \Pr(h|D)$
 - h is hypothesis rainy, fine
 - D is observed data spring|other, cloudy|fine
- ML = maximum likelihood
 - $h_{ml} = \arg \max_h \Pr(D|h)$
- Think of “given” term as function approximator input
- MAP has data input, hypothesis prob output
- ML has h and data as input, *data* prob as output
- ML often uses approximator per hypothesis

From ML to MAP

- MAP is generally preferred
- Bayes Theorem

$$\Pr(h|D) = \frac{\Pr(D|h) \Pr(h)}{\Pr(D)}$$

$$\Pr(h|D) \propto \Pr(D|h) \Pr(h)$$

$$p(h|D) = \Pr(D|h) \Pr(h)$$

- If $\widehat{\Pr(h)}$ is noisy, converting to MAP might hurt
- Hard if we learnt likelihoods instead of probabilities

The problem with ML

- Don't need strict probabilities. Use likelihood $p(D|h)$.
- E.g., log probabilities
- Let $p(D|rain) = f_{rain}(D)$, $p(D|dry) = f_{dry}(D)$

D	cloudy,spring	fine,other
$f_{rain}(D)$	0.1	0.001
$f_{dry}(D)$	0.5	0.9

With uniform priors the conversion to MAP is:

D	cloudy,spring	fine,other
$\Pr(rain D)$	0.99	0.1
$\Pr(dry D)$	0.35	0.65

Most Probable Classification

So far we've sought the most probable *hypothesis* given the data D (i.e., h_{MAP})

Given new instance x , what is its most probable *classification*?

- $h_{MAP}(x)$ is not the most probable classification!

Consider:

- Three possible hypotheses:

$$P(h_1|D) = .4, P(h_2|D) = .3, P(h_3|D) = .3$$

- Given new instance x ,

$$h_1(x) = +, h_2(x) = -, h_3(x) = -$$

- What's most probable classification of x ?

Bayes Optimal Classifier

Bayes optimal classification:

$$\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j | h_i) P(h_i | D)$$

Example:

$$P(h_1 | D) = .4, \quad P(- | h_1) = 0, \quad P(+ | h_1) = 1$$

$$P(h_2 | D) = .3, \quad P(- | h_2) = 1, \quad P(+ | h_2) = 0$$

$$P(h_3 | D) = .3, \quad P(- | h_3) = 1, \quad P(+ | h_3) = 0$$

Bayes Optimal: Example II

Therefore

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4$$

$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6$$

and

$$\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -$$

Gibbs Classifier

Bayes optimal classifier provides best result, but can be expensive if many hypotheses.

Gibbs algorithm:

1. Choose one hypothesis at random, according to $P(h|D)$
2. Use this to classify new instance

Surprising fact: Assume target concepts are drawn at random from H according to priors on H . Then:

$$E[\text{error}_{Gibbs}] \leq 2E[\text{error}_{BayesOptimal}]$$

Suppose correct, uniform prior distribution over H , then

- Pick any hypothesis from V_S , with uniform probability
- Its expected error no worse than twice Bayes optimal

Naive Bayes Classifier

Along with decision trees, neural networks, nearest nbr, one of the most practical learning methods.

When to use

- Moderate or large training set available
- Attributes that describe instances are conditionally independent given classification

Successful applications:

- Diagnosis
- Classifying text documents

Naive Bayes Classifier I

Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2 \dots a_n \rangle$.

Most probable value of $f(x)$ is:

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j | a_1, a_2 \dots a_n)$$

$$v_{MAP} = \arg \max_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)}$$

$$= \arg \max_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j)$$

Naive Bayes Classifier II

Naive Bayes assumption:

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$$

which gives

$$\text{Naive Bayes classifier: } v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

Naive_Bayes_Learn(*examples*)

For each target value v_j

$$\hat{P}(v_j) \leftarrow \text{estimate } P(v_j)$$

For each attribute value a_i of each attribute a

$$\hat{P}(a_i|v_j) \leftarrow \text{estimate } P(a_i|v_j)$$

Classify_New_Instance(x)

$$v_{NB} = \arg \max_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Naive Bayes: Example

Consider *PlayTennis*, and new instance

$\langle \text{Outlk} = \text{sun}, \text{Temp} = \text{cool}, \text{Humid} = \text{high}, \text{Wind} = \text{strong} \rangle$

Want to compute:

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

$$P(y) P(\text{sun}|y) P(\text{cool}|y) P(\text{high}|y) P(\text{strong}|y) = .005$$

$$P(n) P(\text{sun}|n) P(\text{cool}|n) P(\text{high}|n) P(\text{strong}|n) = .021$$

$$\rightarrow v_{NB} = n$$

Naive Bayes: Subtleties I

1. Conditional independence assumption is often violated

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$$

- ...but it works surprisingly well anyway. Note don't need estimated posteriors $\hat{P}(v_j | x)$ to be correct; need only that

$$\arg \max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i | v_j) = \arg \max_{v_j \in V} P(v_j) P(a_1 \dots, a_n | v_j)$$

- see [Domingos & Pazzani, 1996] for analysis
- Naive Bayes posteriors often unrealistically close to 1 or 0

Naive Bayes: Subtleties II

2. what if none of the training instances with target value v_j have attribute value a_i ? Then

$$\hat{P}(a_i|v_j) = 0, \text{ and...}$$

$$\hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = 0$$

Naive Bayes: Subtleties III

Typical solution is Bayesian estimate for $\hat{P}(a_i|v_j)$

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}$$

where

- n is number of training examples for which $v = v_j$,
- n_c number of examples for which $v = v_j$ and $a = a_i$
- p is prior estimate for $\hat{P}(a_i|v_j)$
- m is weight given to prior (i.e. number of “virtual” examples)

Learning to Classify Text

Why?

- Learn which news articles are of interest
- Learn to classify web pages by topic

Naive Bayes is among most effective algorithms

What attributes shall we use to represent text documents??

Learning to Classify Text I

Target concept *Interesting?* : *Document* $\rightarrow \{+, -\}$

1. Represent each document by vector of words
 - one attribute per word position in document
2. Learning: Use training examples to estimate
 - $P(+)$
 - $P(-)$
 - $P(doc|+)$
 - $P(doc|-)$

Learning to Classify Text II

Naive Bayes conditional independence assumption

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

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

One more assumption: $P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

Naive_Bayes_text(*Examples*, *V*)

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

Classify_naive_Bayes_text(Doc)

- $positions \leftarrow$ all word positions in Doc that contain tokens found in $Vocabulary$
- Return v_{NB} , where

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{i \in positions} P(a_i | v_j)$$

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

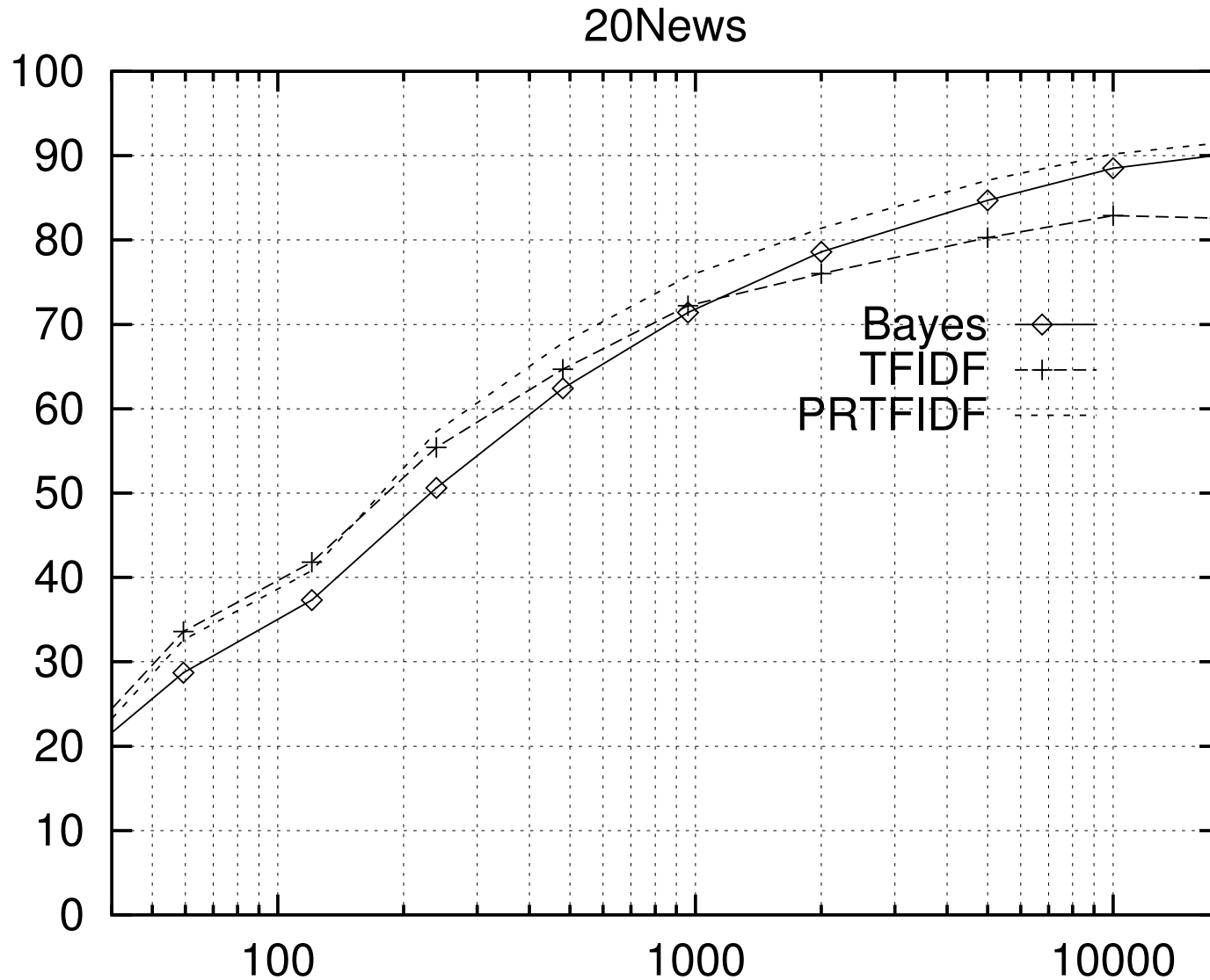
Naive Bayes: 89% classification accuracy

Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!ogicse!uwm.edu
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinion)...
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 Hruday 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

Learning Curve for 20 Newsgroups



Accuracy vs. Training set size (1/3 withheld for test)

Bayesian Belief Networks

Interesting because:

- Naive Bayes assumption of conditional independence too restrictive
 - But it's intractable without some such assumptions...
 - Bayesian Belief networks describe conditional independence among *subsets* of variables
- allows combining prior knowledge about (in)dependencies among variables with observed training data

(also called Bayes Nets)