CS 188 Section Handout: Classification

1 Introduction

In this set of exercises we will experiment with a binary classification problem. The data comprise a training set of *feature* vectors with corresponding *class* labels, and a test set of unlabeled feature vectors which we will attempt to classify with a decision tree, a naive Bayes model, and a perceptron.

Scenario: You are a geek who hates sports. Trying to look cool at a party, you join a lively discussion on professional football and basketball. You have no idea who plays what, but fortunately you have brought your CS188 notes along, and will build some classifiers to determine which sport is being discussed.

Training data: Somehow you come across a pamphlet from the Atlantic Coast Conference Basketball Hall of Fame, as well as an Oakland Raiders team roster. You study these to create the following table:

Sport	*Position*	*Name*	*Height*	*Weight*	*Age*	*College*
Basketball	Guard	Michael Jordan	6'06"	195	43	North Carolina
Basketball	Guard	Vince Carter	6'06"	215	29	North Carolina
Basketball	Guard	Muggsy Bogues	5'03"	135	41	Wake Forest
Basketball	Center	Tim Duncan	6'11"	260	29	Wake Forest
Football	Center	Vince Carter	6'02"	295	23	Oklahoma
Football	Kicker	Tim Duncan	6'00"	215	27	Oklahoma
Football	Kicker	Sebastian Janikowski	6'02"	250	27	Florida State
Football	Guard	Langston Walker	6'08"	345	27	California

Test data: You wish to determine which sport is played by these two subjects of discussion:

Sport	*Position*	*Name*	*Height*	*Weight*	*Age*	*College*
?	Guard	Charlie Ward	6'02"	185	35	Florida State
?	Defensive End	Julius Peppers	6'07"	283	26	North Carolina

2 Naive Bayes Model

In a naive Bayes model, features are conditionally independent given the generating class.

Structure: Draw a graphical model for the naive Bayes sports classifier.

Solution: A tree with root node, variable *Sport*, whose children are the evidenced feature variables.

Estimation: Compute empirical distributions with the maximum-likelihood estimate:

$\hat{P}(S)$

B	1/2
F	1/2

$\hat{P}(P \mid S)$

G	B	3/4
C	B	1/4
G	F	1/4
C	F	1/4
K	F	1/2

$\hat{P}(N \mid S)$

MJ	B	1/4
VC	B	1/4
TD	B	1/4
MB	B	1/4
VC	F	1/4
TD	F	1/4
LW	F	1/4
SJ	F	1/4

$\hat{P}(H > 6\text{'}03\text{"} \mid S)$

True	B	3/4
False	B	1/4
True	F	1/4
False	F	3/4

$\hat{P}(W > 200 \mid S)$

True	B	1/2
False	B	1/2
True	F	1
False	F	0

$\hat{P}(A > 28 \mid S)$

True	B	1
False	B	0
True	F	0
False	F	1

$\hat{P}(C \mid S)$

NC	B	1/2
WF	B	1/2
O	F	1/2
FS	F	1/4
C	F	1/4

Smoothing: Apply *add-one* (Laplace) smoothing to distributions, allow possible unknown (UNK) values:

$\hat{P}(S)$

B	1/2
F	1/2

$\hat{P}(P \mid S)$

G	B	4/7
C	B	2/7
UNK	B	1/7
G	F	1/4
C	F	1/4
K	F	5/8
UNK	F	1/8

$\hat{P}(N \mid S)$

MJ	B	2/9
VC	B	2/9
TD	B	2/9
MB	B	2/9
UNK	B	1/9
VC	F	2/9
TD	F	2/9
LW	F	2/9
SJ	F	2/9
UNK	F	1/9

$\hat{P}(H > 6\text{'}03\text{"} \mid S)$

True	B	2/3
False	B	1/3
True	F	1/3
False	F	2/3

$\hat{P}(W > 200 \mid S)$

True	B	1/2
False	B	1/2
True	F	5/6
False	F	1/6

$\hat{P}(A > 28 \mid S)$

True	B	5/6
False	B	1/6
True	F	1/6
False	F	5/6

$\hat{P}(C \mid S)$

NC	B	3/7
WF	B	3/7
UNK	B	1/7
O	F	3/8
FS	F	1/4
C	F	1/4
UNK	F	1/8

Inference: For the first test example, compute the posterior probability of *Sport* using the smoothed and unsmoothed empirical probability estimates.

Solution: By conditional independence:

$$P(Sport, Position, Name, Height > 6\text{'}03\text{"}, Weight > 200, Age > 28, College) =$$
$$P(Sport)P(Position|Sport)P(Name|Sport)P(Height > 6\text{'}03\text{"}|Sport)P(Weight > 200|Sport)P(Age > 28|Sport)P(College|Sport)$$

With unsmoothed probabilities, both test examples have joint probability of zero for either class because values unseen in training are given no empirical probability mass, e.g. $P(Charlie\ Ward \mid Sport) = 0$.

Using the smoothed probabilities and UNK for values unseen in training:

$$P(B)P(G|B)P(CW|B)P(H < 6\text{'}03\text{"}|B)P(W < 200|B)P(A > 28|B)P(FS|B) = (1/2)(4/7)(1/9)(1/3)(1/2)(5/6)(1/7) = \frac{5}{7938}$$

$$P(F)P(G|F)P(CW|F)P(H < 6\text{'}03\text{"}|F)P(W < 200|F)P(A > 28|F)P(FS|F) = (1/2)(1/4)(1/9)(2/3)(1/6)(1/6)(1/4) = \frac{1}{15552}$$

$$P(Basketball \mid Guard, Charlie\ Ward, Height < 6\text{'}03\text{"}, Weight < 200, Age > 28, Florida\ State) = \frac{\frac{5}{7938}}{\frac{5}{7938} + \frac{1}{15552}} \approx 0.91$$

Extra: Charlie Ward and Julius Peppers were athletes who each competed in both football *and* basketball during college. If you trained these classifiers correctly, you can predict what professional sports they chose.

Solution: In 1993, Charlie Ward led his basketball team to the Elite Eight, led his football team to the national championship, and won the Heisman Trophy (nation's best college football player). Ward passed up the NFL (or rather, scouts claimed he was too small for quarterback – this was before the Doug Flutie phenomenon) but he was drafted in the first round by the New York Knicks, where he had a good career before finishing in San Antonio and Houston. He is now an assistant coach for the Rockets.

Julius Frazier Peppers was named after two basketball legends (Julius Erving and Walt Frazier). He attended UNC on a football scholarship, but was able walk on to the basketball team, where he had a 64% field goal percentage. Though it was his preferred sport and he probably could have made it in the NBA, Peppers quit basketball and chose to enter the NFL, where he has been pretty awesome so far.

Other two-sport legends include a pair of NFL tight ends. Cal's own Tony Gonzalez went to the Sweet 16 in basketball before joining the Kansas City Chiefs. Antonio Gates led the Kent State Golden Flashes to the Elite 8; he did not play football in college but was signed undrafted by the San Diego Chargers, where he has been an unbelievable success.