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 Decision Trees

A decision tree classifies a feature vector by asking "questions" of its *attributes*, and returns a predicted "decision" based on a structure learned to represent the training data.

Entropy: What is the information content of the initial training distribution of *Sport*?

Solution: Initially, 4 basketball and 4 football players, so a $\langle \frac{1}{2}, \frac{1}{2} \rangle$ distribution, with 1 bit of information.

Information gain: Calculate the information gain for the decision stumps created by first splitting on $Position$, $Name$, or $College$? Which of these compact decision trees perfectly classifies the training data?

Solution: Calculate the expected entropy, weighted by proportion of examples in each split set. Splitting on $College$ is best, resulting in an expected entropy of 0 (gain of 1).

Feature Engineering: Decision trees can represent any function of discrete attribute variables. For continuous variables ($Height$, $Weight$, and Age), show how to reasonably cast these as Boolean variables.

Solution: Use an inequality relation, $Attribute > a$, where a is a split point chosen to give the highest information gain (AIMA pg. 664). E.g., an initial split on $Age > 28$ will perfectly classify the training data.

Expressiveness: Draw a few decision trees that each correctly classify the training data, and show how their predictions vary on the test set.

Solution: There are many complete decision trees of various depths. When testing the second example, there is a problem because $Position$ has an attribute value that is missing from the trained decision tree. See AIMA pg. 663 and Ex. 18.12 for discussion of this issue, and one way to handle it.

3 Perceptrons

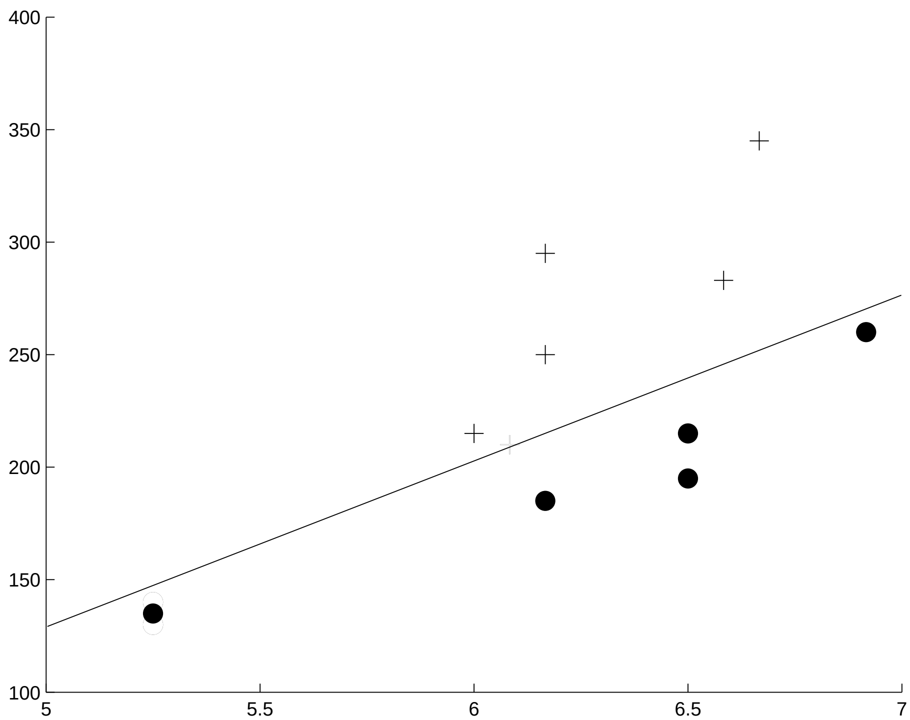
A perceptron can be applied to input features with continuous values. The weights of the perceptron define hyperplanes that represent decision boundaries for classification. Assume we are only using the input features $f_1 = Height$ and $f_2 = Weight$, along with a bias $f_0 = 1$.

Structure: Draw the networks for a binary and a multi-class perceptron sports classifier.

Solution: Binary: three input nodes, one output node, three weights, and a threshhold activation function. Multi-class: three input nodes, two output nodes, six weights, and argmax decision function.

Linear separability: A binary or two-class perceptron can only represent linearly separable functions. Plot the training data below to determine if there exists a linear separator.

Solution: Green dots are basketball, red crosses are football, and cyan is used for the test data.



Learning: Without using the learning algorithm for perceptron weight updates, can you use the plot above to guess values for the weights of a binary perceptron that correctly classifies all the training data?

Solution: Fit a line that separates the basketball examples from the football examples with some margin. Because this line represents the decision boundary where the binary perceptron's output is zero, the equation for this line in the form $0 = c + af_1 + bf_2$ corresponds to the relation among weights $0 = w_0 + w_1 f_1 + w_2 f_2$ (equivalently, the hyperplane $0 = \mathbf{w} \cdot \mathbf{f}$). Note that the bias weight allows linear separators that don't necessarily pass through the origin.

4 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.