Problem Set 2 — PMR3508

Prof. Fabio Cozman — 2nd semester 2018

1) What is:

a) An artificial neural network?

b) A perceptron?

c) The function encoded by a perceptron?

d) The learning rule for the perceptron?

e) The disadvantages of the perceptron as classifier?

f) A multilayer perceptron?

g) A hidden layer?

h) The recall of a multilayer perceptron?

i) The sigmoid logistic function?

j) The backpropagation algorithm?

k) Deep learning with multilayer perceptrons?

l) Autoencoders?

m) ReLU activation functions?

2) What is:

a) The bias-variance trade-off in regression problems?

b) A covariate in a regression problem? And a response variable?

c) One-hot encoding?

d) Linear regression? Simple regression?

e) Residual sum of squares (RSS)?

f) The meaning of a test for $\beta_1 = 0$ in simple regression?

g) The meaning of the R^2 statistic? What is it used for? (Recall: $R^2 = 1 - \text{RSS}/\sum_j (y_j - \hat{y})^2$.)

h) Feature selection? Why is it often useful?

i) Forward stepwise regression? Backward stepwise regression?

j) The application of the Akaike Information Criterion (AIC)? (Recall: AIC $= L_S - |S|$.)

k) The application of the Bayesian Information Criterion (BIC)? (Recall: BIC $= L_S - (|S|/2) \log_2 N$.)

l) The application of the Mallow's C_p score? (Recall: $C_p = 2|S|\hat{\sigma}^2 + \sum_{j=1}^{N} (\hat{y}_j - y_j)^2$, where $\hat{\sigma}^2$ is the estimate of variance of residuals.)

m) Ridge regression?

n) The effect of ridge reresssion on variance? And on bias?

o) Lasso?

p) The effect of lasso on variance? And on bias?

q) The main advantage of lasso?

3) What is:

a) Logistic regression?

b) The likelihood of the training set in logistic regression?

c) The expression for $\mathbb{P}(Y = 1|\text{data})$ in logistic regression?

d) Tbe expression for the log-odds of the class variable in logistic regression?

e) The function logit?

f) The purpose of the IRLS algorithm?

g) Linear discriminant analysis (LDA)?

h) The assumptions in LDA?

i) The likelihood of the training set in LDA?

j) The classification boundary of LDA, given estimates for all parameters in the prior and likelihood?

4) What is:

a) A classification tree?

b) A regression tree?

c) A split in a classification/regression tree?

d) The main loop in C4.5?

e) The application of entropy/information gain/gain ratio in C4.5?

f) The application of the Gini index in CART?

g) The advantages of a classification/regression tree?

h) The disadvantages of a classification/regression tree?

5) What is:

a) The main ideas in boosting?

b) A weak classifier in boosting?

c) The AdaBoost algorithm?

d) A stump?

e) Gradient boosting?

6) What is:

a) A linearly separable problem?

b) The margin of a separating hyperplane in a linearly separable problem?

c) A support vector for a separating hyperplane in a linearly separable problem?

d) A soft margin?

e) A support vector machine?

7) What is:

a) Unsupervised learning?

b) Representation learning?

c) Clustering?

d) Principal Component Analysis (PCA)?

e) How is each vector of loadings defined in PCA?

f) An eigenvalue? An eigenvector?

g) The relationship between PCA and eigenvalues/eigenvectors?

h) K-means?

i) The centroid of a cluster?

j) The inputs to K-means? How do they affect the algorithm?

k) The possible ways to start up K-means?

l) The possible ways to stop K-means?

m) The advantages of K-means?

n) The disadivantages of K-means?

o) Hierarchical clustering?

p) Dendrogram?

q) Agglomerative (bottom-up) clustering?

r) The possible ways to compute the distance between clusters in agglomerative clustering?

8) Suppose we have two integer-valued features X_1 and X_2, and each pair (x_1, x_2) may be observed with identical probability for $x_1 \in \{1, 2, \ldots, 10\}$ and $x_2 \in \{1, 2, \ldots, 10\}$. Moreover, suppose that we have three classifiers g_1, g_2 and g_3 that classify each point with label 0 or 1, such that

- g_1 makes no mistake for $x_1 > 3$; otherwise, g_1 makes a mistake with probability 1/6.
- g_2 makes no mistake for $x_2 > 2$; otherwise, g_2 produces an incorrect output.
- g_3 makes no mistake for $x_1 < 7$; otherwise, g_3 makes a mistake with probability 1/4.

The mistakes made by the classifiers are independent. We are interested in boosting these classifiers by defining a classifier g that works by *majority rule*: for a given input (x_1, x_2), the output of g is the output that is most frequent in the set $\{g_1(x_1, x_2), g_2(x_1, x_2), g_3(x_1, x_2)\}$.

What is the error rate for g_1, g_2, g_3, and g? Is it worth boosting the three initial classifiers by majority rule?

9) Consider a perceptron with three inputs, X_1, X_2 and X_3, and a continuous output Y that is produced by a logistic activation function $f(x_1, x_2, x_3)$ of the weighted sum of inputs W:

$$Y = \frac{1}{1 + \exp(-(1/3)W)},$$

where W is computed with weights 1, 2, and 4 for inputs X_1, X_2 and X_3 respectively.

What is the output Y for inputs $(x_1, x_2, x_3) = (-3, 10, -2)$?

What is the output Y if the logistic activation function is replaced by a ReLU activation function?

10) Consider a training set as follows:

X_1	X_2	Y
1	2	0
1	1	1
2	2	0
2	2	0
2	2	1
1	1	1

X_1	X_2	Y
1	1	0
1	2	1
1	1	0
2	1	1
2	2	0
2	1	0

Suppose we want to build a stump, and we need to decide whether to split using the Gini index (CART algorithm).

What are the necessary computations to decide on the split? What is the result?

What are the necessary computations if we decide to split on gain ration (C4.5 algorithm)? What is the result?

11) Suppose we have two features X_1 and X_2 that were normalized to have zero mean. We just have two observations:

$$x_1 = 4, x_2 = 3, \qquad \text{and} \qquad x_1 = -4, x_2 = -3.$$

We then find the first principal component, Z_1.

What is the expression for Z_1 as a function of X_1 and X_2?

12) Consider a continuous feature X_1 and a binary class variable Y, and three measurements (X_1, Y): first, $(-1, 1)$; second, $(0, -1)$; third, $(1, 1)$.

This problem is not linearly separable. (Why?)

Consider an enlarged vector of features (X_1, X_1^2). Determine whether this vector leads to a linearly separable problem; if so, find a SVM for the problem and compute its margin in the enlarged space.

13) Suppose you have a covariate X and a response variable Y, and observations $(x_1, y_1), \ldots, (x_N, y_N)$. Suppose also that you want to find the function

$$Y = \beta X$$

that minimizes the residual sum of squares (that is, you must find β).

Write down the residual sum of squares as a polynomial

$$a\beta^2 + b\beta + c.$$

What is a, b, and c?

What is the value of β that minimizes the residual sum of squares (that is, that minimizes this polynomial)?