Brown University
DATA 1010
Fall 2018: Practice Midterm III
Samuel S. Watson

Name:

You will have three hours to complete the exam, which consists of 36 questions. Among the first 24 questions, you should only solve problems for standards for which you want to improve your medal from the second exam.

No calculators or other materials are allowed, except the provided reference sheets.

You are responsible for explaining your answer to **every** *question. Your explanations do not have to be any longer than necessary to convince the reader that your answer is correct.*

For questions with a final answer box, please write your answer as clearly as possible and strictly in accordance with the format specified in the problem statement. Do not write anything else in the answer box. Your answers will be grouped by Gradescope's AI, so following these instructions will make the grading process much smoother.

I verify that I have read the instructions and will abide by the rules of the exam: ______________________________

Problem 1 **[STATLEARN]**

Give an example which shows that a simple linear regression model can overfit the data. Give some ideas for how to mitigate the overfitting.

Solution

Problem 2 [LRC]

Show that the Bayes classifier is the classifier which minimizes the misclassification probability.

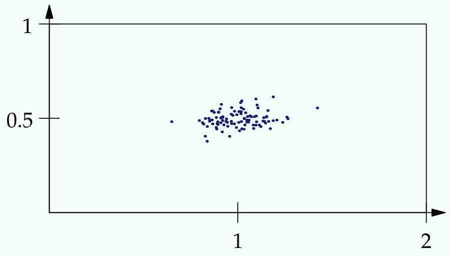
For simplicity, you may assume the context of a binary classification problem with a discrete sample space.

Solution

Problem 3 [KDE]

(a) Suppose that $f_\lambda(x, y)$ is the bandwidth-λ kernel density estimator associated with the set of samples shown in the figure (based on the tri-cube weight function).

Estimate the value of λ such that the points (x, y) for which $f_\lambda(x, y) = 0$ make up approximately half of the rectangle by area.



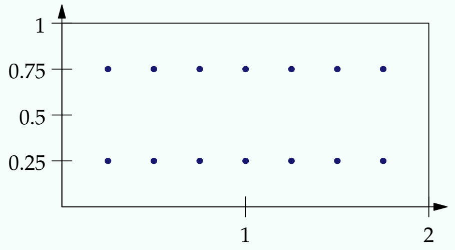
(b) Consider a set of points $\{(x_i, y_i)\}_{i=1}^n$ in $\mathbb{R}^2$ and a positive value of λ. Suppose that the vertical line $x = a$ passes through the three sample points $(x_1, y_1), (x_2, y_2)$, and (x_3, y_3), and that no other sample points have an x value within λ of a. Find the value of $r_\lambda(a)$ (where r_λ is the Nadaraya-Watson estimator associated with the samples).

Solution

Final answer:

Problem 4 **[LR]**

Find the residual sum of squares for the line of best fit for the samples shown.



Solution

Final answer:

Problem 5 **[LOGIST]**

Consider a binary classification problem for which there exists a hyperplane separating the classes. What goes wrong if you try to apply logistic regression?

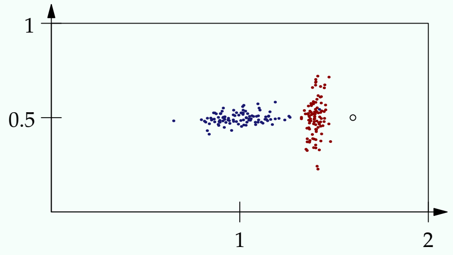
Solution

Problem 6 [QDA]

Select which of the two statements is correct (given that one of them is correct), and explain why the two classifiers behave differently.

(a) The point marked with a hollow circle is classified as blue by a QDA classifier and red by a kernel density classifier.

(b) The point marked with a hollow circle is classified as red by a QDA classifier and blue by a kernel density classifier.

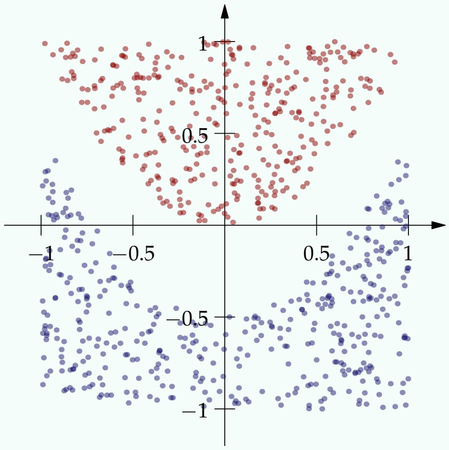


Solution

Final answer:

Problem 7 [SVM]

Find a map from the plane to some other Euclidean space such that hard-margin SVM could, after applying the map, be used for classification problem shown in the figure.



Solution

Problem 8 [DT]

Are decision tree classifiers scale sensitive? In other words, if a feature is scaled by the same constant factor for all observations, do we end up with a different trained decision tree classifier?

Solution

Problem 9 **[ENSEMBLE]**

For each of the following assertions, determine whether it is true or false.

(a) If the individual models which make up an ensemble classifier do not have very high accuracy, then the ensemble classifier will also have pretty low accuracy.

(b) An ensembled regression function always has lower loss than its constitutent models individually.

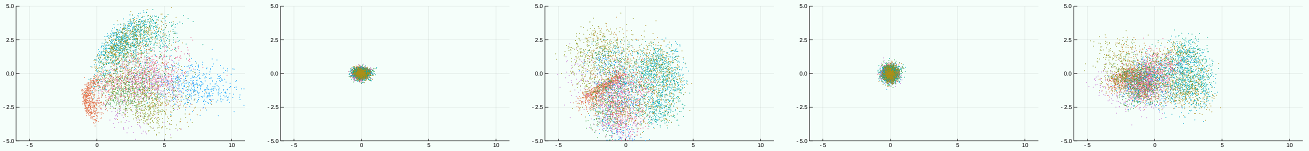
(c) The constituent models must be independent for ensembling to work.

(d) Bagging is analogous to a referendum, while gradient boosting is like getting better over time incrementally, with each step taken to make the best improvement we can given our constraints.

Solution

Problem 10 [DR]

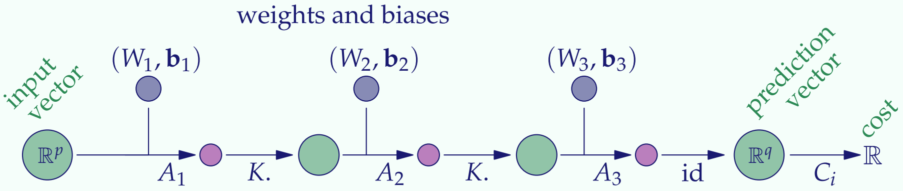
The first graph below shows the dot product of each of the first 5000 (de-meaned) vectors in the MNIST training set with the first principal component and the second principal component. The remaining graphs are similar, but using different pairs of principal components. One uses the second and third, one uses the second and tenth, one uses the 80th and 81st, and one uses the 80th and 120th. Figure out which is which.



Solution

Problem 11 **[NN]**

Suppose that weights and biases have been chosen for the neural network shown, and that a vector has been forward propagated through the network. Suppose that the vectors recorded at the purple nodes are $[1, -4, 2]$, $[6, 3]$, and $[9, 7, -4, -1, 5]$.



(a) What is the architecture of this neural net?

(b) What vector is recorded at the second green node (the one between A_1 and A_2)?

(c) Now suppose that we are in the midst of the backpropagation process, and we have just determined that the derivative of the cost with respect to the vector in the second purple node is equal to $[-3, -4]'$. Calculate the derivative of the cost with respect to the matrix W_2.

Solution

Problem 12 [FREQBAYES]

(a) Explain why conjugate priors are an exclusively Bayesian statistics topic (in other words, explain why they are not useful/meaningful in the frequentist framework).

(b) Outline a strategy for computing a Bayesian point estimate, supposing that our prior distribution is not a conjugate prior for the problem at hand.

Solution