PRACTICE QUESTIONS

- If you train a linear regression estimator with only half the data, its bias is smaller.

- Suppose you are given a dataset of cellular images from patients with and without cancer. If you are required to train a classifier that predicts the probability that the patient has cancer, you would prefer to use Decision trees over logistic regression.

- Suppose the dataset in the previous question had 900 cancer-free images and 100 images from cancer patients. If I train a classifier which achieves 85% accuracy on this dataset, it is it a good classifier.

- A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set.

- Assume m is the minimum number of training examples sufficient to guarantee that with probability $1 - \delta$ a consistent learner using hypothesis space H will output a hypothesis with true error at worst ε. Then a second learner that uses hypothesis space H' will require 2m training examples (to make the same guarantee) if $|H'| = 2|H|$.

- Which of the following classifiers can perfectly classify the XOR data:

+	-
-	+

  (a) Decision Tree, (b) Logistic Regression, (c) Gaussian Naive Bayes

- When the feature space is larger, overfitting is less likely.

- Non-parametric models are usually more efficient than parametric models in terms of model storage.

- Boosting decision stumps can result in a quadratic decision boundary.

- Suppose you wish to predict age of a person from his/her brain scan using regression, but you only have 10 subjects and each subject is represented by the brain activity at 20,000 regions in the brain. You would prefer to use least squares regression instead of ridge regression.

- When doing kernel regression on a memory-constrained device, you should prefer to use a box kernel instead of a Gaussian kernel. ?

- To predict the chance that Steelers football team will win the Super Bowl Championship next year, you should prefer to use logistic regression instead of decision trees.

- The kmeans algorithm finds the global optimum of the kmeans cost function.

- Unlike the k-means objective, it is computationally feasible to find the optimal parameters for Gaussian mixture models exactly since the cluster assignments in GMM are soft. ?

PRACTICE QUESTION ANSWERS

- If you train a linear regression estimator with only half the data, its bias is smaller.

SOLUTION: FALSE. Bias depends on the model you use (in this case linear regression) and not on the number of training data.

- Suppose you are given a dataset of cellular images from patients with and without cancer. If you are required to train a classifier that predicts the probability that the patient has cancer, you would prefer to use Decision trees over logistic regression.

SOLUTION: FALSE. Decision trees only provide a label estimate, whereas logistic regression provides the probability of a label (patient has cancer) for a given input (cellular image).

- Suppose the dataset in the previous question had 900 cancer-free images and 100 images from cancer patients. If I train a classifier which achieves 85% accuracy on this dataset, it is it a good classifier.

⋆ SOLUTION: FALSE. This is not a good accuracy on this dataset, since a classifier that outputs ”cancer-free” for all input images will have better accuracy (90%).

- A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set.

⋆ SOLUTION: FALSE. The second classifier has better test accuracy which reflects the true accuracy, whereas the first classifier is overfitting.

- Assume m is the minimum number of training examples sufficient to guarantee that with probability $1 - \delta$ a consistent learner (i.e. a learner that correctly classifies the training data) using hypothesis space H will output a hypothesis with true error at worst ε. Then a second learner that uses hypothesis space H' will require 2m training examples (to make the same guarantee) if $|H'| = 2|H|$.

⋆ SOLUTION: FALSE. Minimum number of training examples sufficient to make an (ε,δ)-PAC guarantee depends logarithmically on hypothesis class size ($\ln|H|$) and not linearly.

- Which of the following classifiers can perfectly classify the XOR data:

+	-
-	+

(a) Decision Tree (b) Logistic Regression (c) Gaussian Naive Bayes

⋆ SOLUTION: Decision Tree only. Decision trees of depth 2 which first splits on X_1 and then on X_2 wil perfectly classify it. Logistic regression leads to linear decision boundaries, hence cannot classify this data perfectly. Due to conditional independence requirement, it is not possible to fit a Gaussian that peaks at the labels of only one class and has no covariance between features, so Gaussian Naive Bayes cannot classify this data perfectly.

- When the feature space is larger, overfitting is less likely.

⋆ SOLUTION: False. The more the number of features, the higher the complexity of the model and hence greater its ability to overfi the training data.

- Non-parametric models are usually more efficient than parametric models in terms of model storage.

⋆ SOLUTION: False. Non-parametric models either need to look at the entire dataset to predict the label of test points or require the number of parameters to scale with the dataset size, hence require more storage.

- Boosting decision stumps can result in a quadratic decision boundary.

⋆ SOLUTION: False. The sign of a finite linear combination of decision stumps always results in a piecewise linear decision boundary.

e.g.

d2 boost

- Suppose you wish to predict age of a person from his/her brain scan using regression, but you only have 10 subjects and each subject is represented by the brain activity at 20,000 regions in the brain. You would prefer to use least squares regression instead of ridge regression.

⋆ SOLUTION: False. When the number of datapoints (subjects) is less than number of features, the least squares solution needs to be regularized to prevent overfitting, hence we prefer ridge regression.

- When doing kernel regression on a memory-constrained device, you should prefer to use a box kernel instead of a Gaussian kernel.

⋆ SOLUTION: True. A box kernel only uses a few data points for prediction and hence does not need to load the entire dataset into memory unlike Gaussian kernel which assigns non-zero weight to all training data points.

- To predict the chance that Steelers football team will win the Super Bowl Championship next year, you should prefer to use logistic regression instead of decision trees.

⋆ SOLUTION: True. Logistic regression will characterize the probability (chance) of label being win or loss, whereas decision tree will simply output the decision (win or loss).

- The kmeans algorithm finds the global optimum of the kmeans cost function.

⋆ SOLUTION: False. The kmeans cost function is non-convex and the algorithm is only guaranteed
to converge to a local optimum.

- Unlike the k-means objective, it is computationally feasible to find the optimal parameters for Gaussian mixture models exactly since the cluster assignments in GMM are “soft."

⋆ SOLUTION: False. The maximum likelihood optimization for GMM is still non-convex.