Problem 1. (12 points)

(a) Consider a data set with N samples $(x^{(i)}, y^{(i)})$. We create a model to predict y: $y^{(i)} = \omega_0 + \omega_1 x^{(i)} + \epsilon_i$, where ϵ_i is random noise. We train our model to minimize MSE, so that

$$(\hat{\omega}_0, \hat{\omega}_1) = \text{argmin} \sum_{i=1}^{N} (y^{(i)} - \omega_0 - \omega_1 x^{(i)})^2.$$

(1) True or False (circle one): We can optimize ω_0 and ω_1 in closed form, using matrix inversion.

(2) Which one of the following will be true after training our linear regression model? (circle one)

(i) $\sum_{i=1}^{N} (y^{(i)} - \omega_0 - \omega_1 x^{(i)}) y^{(i)} = 0$

(ii) $\sum_{i=1}^{N} (y^{(i)} - \omega_0 - \omega_1 x^{(i)}) (x^{(i)})^2 = 0$

(iii) $\sum_{i=1}^{N} (y^{(i)} - \omega_0 - \omega_1 x^{(i)}) (x^{(i)}) = 0$

(iv) $\sum_{i=1}^{N} (y^{(i)} - \omega_0 - \omega_1 x^{(i)})^2 = 0$

(b) Suppose we believe our model to be overfitting. We decide to increase the number of training data used by our learner. Choose one answer for each part.

(1) Training error will most likely [increase decrease stay the same]

(2) Test error will most likely [increase decrease stay the same]

Now suppose we instead believe our model to be underfitting. We again increase the number of training data used by our learner. Choose one answer for each part.

(1) Training error will most likely [increase decrease stay the same]

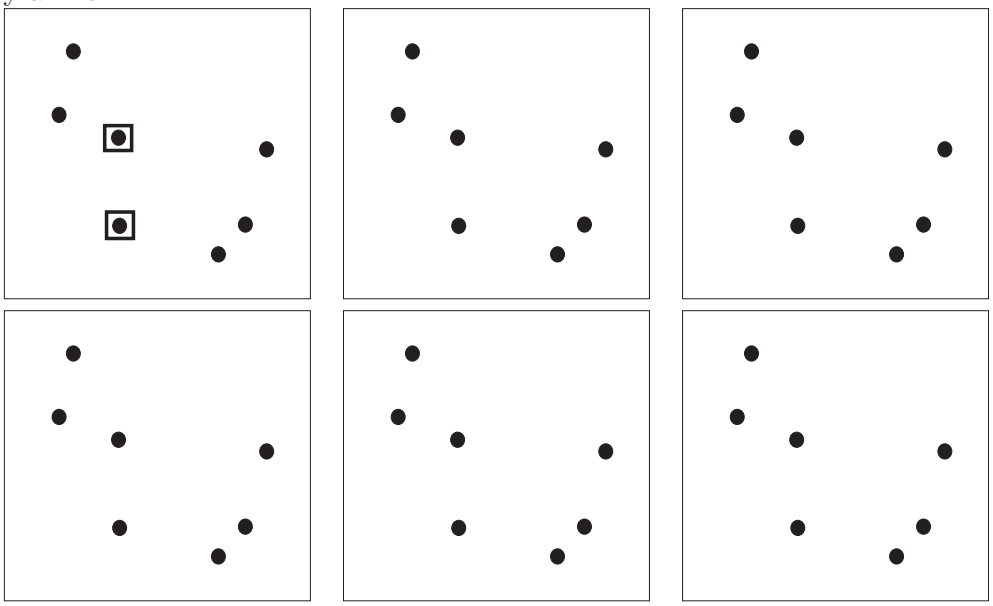
(2) Test error will most likely [increase decrease stay the same]

Problem 2: (18 points) Clustering

Consider the two-dimensional data points plotted in each panel. In this problem, we will cluster these data using two different algorithms, where each panel is used to show an iteration or step of the algorithm.

k-means

(a) Starting from the two cluster centers indicated by squares, perform k-means clustering on the data points. In each panel, indicate (somehow) the data assignment, and in the next panel show the new cluster centers. Stop when converged, or after 6 steps, whichever is first. It may be helpful to recall from our nearest-neighbor classifier that the set of points nearer to A than B is separated by a line.

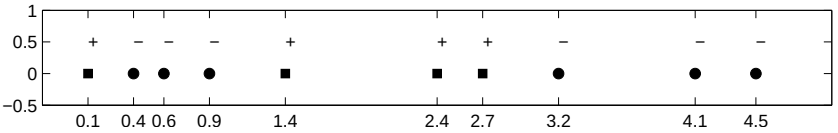


(b) Write down the cost function optimized by the k-means algorithm, and explain your notation.

(c) Give an advantage of hierarchical agglomerative clustering over k-means and an advantage of k-means over hierarchical clustering.

Problem 3: (16 points) Classification in One Dimension

We observe a collection of training data with one feature, "x" and a class label $c \in \{-, +\}$, shown here; class + is indicated by squares and − by circles, and also labelled with text. Answer each of the following questions. Express error rates as the fraction of data points incorrectly classified.



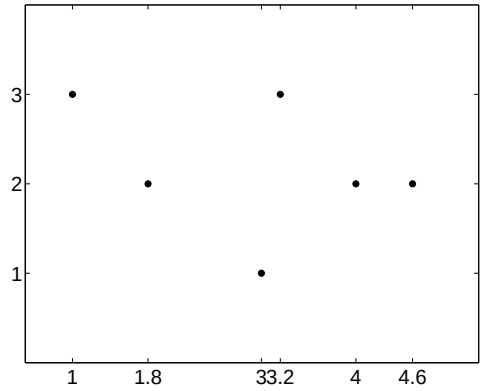
(a) What is the best training error rate we can achieve on these data from a Gaussian model Bayes classifier with *equal* variances for the two classes? Explain briefly (sketch + 1-2 sentences): how is it achieved and why is it the best?

(b) What is the best training error we can achieve from a Gaussian Bayes classifier with *arbitrary* variances for the two classes? Explain briefly how it's achieved and why it's the best.

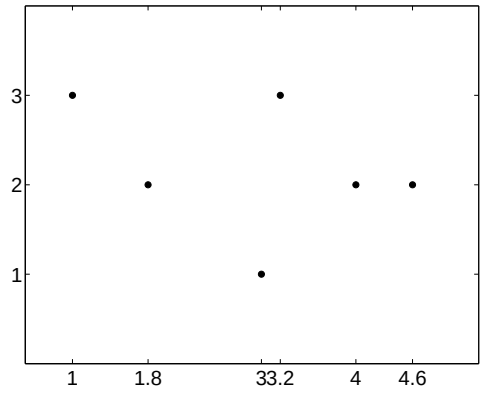
l

Problem 4: (12 points) kNN Regression

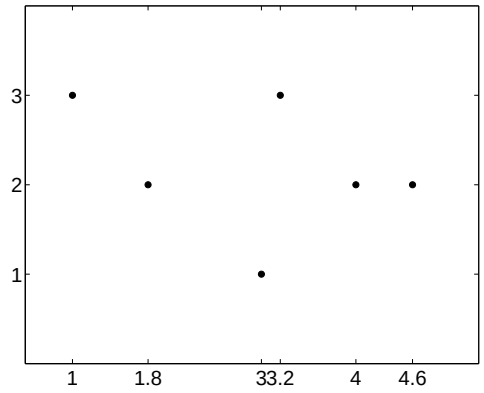
Consider a regression problem for predicting the data shown at left using your k-nearest neighbor regression algorithm from the homework. Under each of the following scenarios, (a) sketch the regression function when trained on all the data; (b) compute its resulting training error (MSE). (If you like you may leave an arithmetic expression, i.e., leave values as for example "$(.6)^2$")



(a) $k = 1$



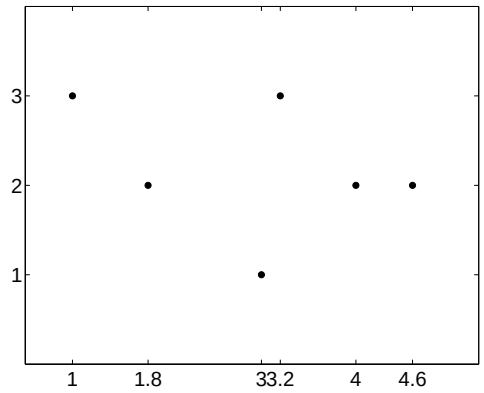
(b) $k = 2$



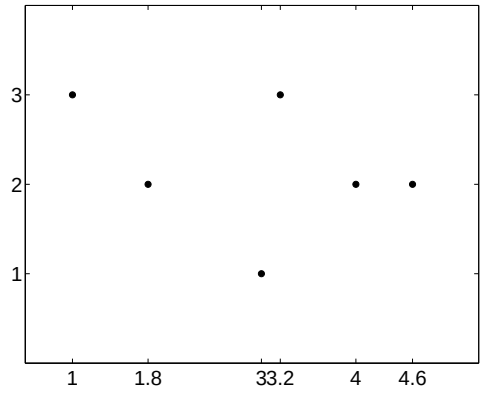
(c) $k = 5$

Problem 5: (10 points) Cross-validation Errors

Using the same data, compute the *leave-one-out cross-validation* MSE error rate (i.e., the 6-fold cross-validation error rate). Use the figures to sketch the predicted value at each training point *when it is left out*. Again, you may leave un-evaluated arithmetic expressions.



(a) $k = 1$



(b) $k = 5$

Problem 6: (16 points) Bayes Classifiers

In this problem you will use Bayes Rule, $p(y|x) = p(x|y)p(y)/p(x)$ to perform classification. Suppose we observe the following training data, with three binary features $x_1, \ldots, x_3$ and a binary class y:

x_1	x_2	x_3	y
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	1
0	0	1	1
0	1	0	1
1	1	0	1

Learn to predict y using a naïve Bayes classifier; **show your work**.

(a) After learning the model, what is the predicted probability $p(y = 0|x_1 = 0, x_2 = 1, x_3 = 0)$?

(b) Suppose that we *only* observe $x_1 = 0$. What is the predicted probability $p(y = 0|x_1 = 0)$?

For the next two parts, learn a *joint* Bayes classifier.

(a) What is the predicted probability $p(y = 0|x_1 = 0, x_2 = 1, x_3 = 0)$?

(b) What is $p(y = 0|x_1 = 0)$?

Problem 7: (16 points) Regression

We have a dataset consisting of M examples, each with one feature x and a real-valued target y.

(a) Suppose we use linear regression to fit the data, using D polynomial features. From our full database, we choose M_1 data points (at random) to be a training data set, and M_2 of the remaining data to be a validation (test) set.
Suppose we begin to increase the set of available features by, for example, taking polynomials of x. What do you expect to happen to the mean squared error (1) on the training set and (2) on the test set? Sketch and explain in 1-2 sentences.

(b) Now suppose that we train a *non-linear* regression model, where our prediction is

$$\hat{y}(x) = \exp(\theta\, x)$$

with a single, scalar parameter θ.

(1) Write down the formula for the mean squared error on the training data.

(2) Suppose that we train our model to minimize the (training) MSE. Which of the following will be true of θ? (Circle one)

i. $\sum_i y^{(i)} = \sum_i x^{(i)} \exp(\theta x^{(i)})$

ii. $\sum_i y^{(i)} \exp(\theta x^{(i)}) = \sum_i x^{(i)} \exp(\theta x^{(i)})$

iii. $\sum_i y^{(i)} x^{(i)} \exp(\theta x^{(i)}) = \sum_i x^{(i)} \exp(\theta x^{(i)})$

iv. $\sum_i y^{(i)} x^{(i)} \exp(\theta x^{(i)}) = \sum_i x^{(i)} \exp(2\theta x^{(i)})$