

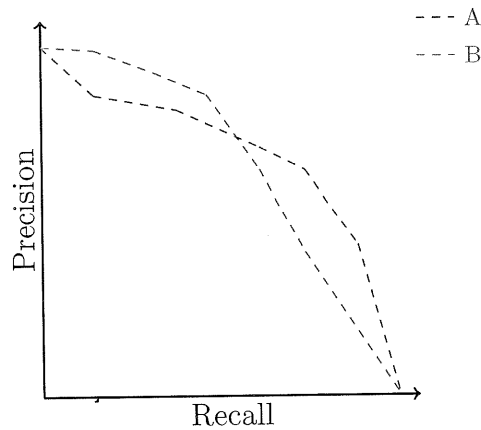
---

## True/False

1.  T  F Linear regression is a useful model to make predictions, but it is limited by the fact that we are unable to interpret the model to make inferences about the relationships between features and the output.
2.  T  F When given the choice of two models, the one that has smaller training error will always have smaller true error.
3.  T  F We expect a model with high variance to generalize better than a model with high bias.
4.  T  F When we use the same model complexity on a smaller dataset, overfitting is more likely.
5.  T  F In machine learning, bias is always a bigger source of error than variance.
6.  T  F Given an infinite amount of noiseless training data, we expect the training error for decision stumps to go to 0.
7.  T  F As the number of iterations goes to infinity, boosting is guaranteed to reach zero training error. *(Depends if assumptions are true, would accept False)*
8.  T  F Increasing  $k$  in  $k$ -NN increases bias and decreases variance.
9.  T  F To determine the best value of  $k$  for  $k$ -means, it's sufficient to run the  $k$ -means algorithm once for each value of  $k$  you want to try.
10.  T  F Increasing the number of recommended items is more likely to increase the recall than decrease it.
11.  T  F With a large dataset, nearest neighbors is more efficient at test time than logistic regression.
12.  T  F  $k$ -means converges to a global optimum for the heterogeneity objective.
13.  T  F To find the best set of coefficients for logistic regression, we use gradient descent to minimize the number of examples misclassified.

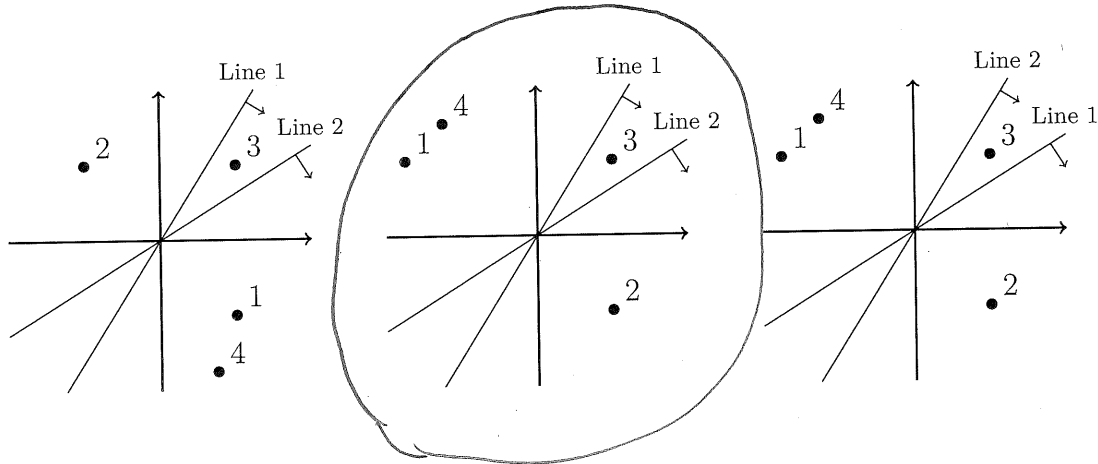
## Multiple Choice

1. Select all the following reasons why we would use LASSO over Ridge:
  - It can help us identify which features are important
  - It is faster to learn the weights for LASSO than for Ridge
  - LASSO usually achieves lower generalization error than Ridge
  - If there are many features, the model learned using LASSO can make predictions more efficiently.
2. Which of the following are symptoms of a logistic regression model being overfit? Select all that apply
  - Large estimated coefficients
  - Good generalization to unseen data
  - Simple decision boundary
  - Complex decision boundary
  - Overconfident predictions of class probabilities
3. For the following precision-recall curves for varying classification thresholds on different models A and B, which is the best model?
  - Model A
  - Model B
  - Depends on the situation



4. Suppose we had the following bins for locality sensitive hashing. The small arrows perpendicular to the lines indicate which side of the boundary is classified as +1. Circle the graph that represents the data in the table.

Bin Index	0 0	0 1	1 0	1 1
Point Labels	{1,4}		{3}	{2}



## Short Response

1. Give examples of 3 classification models we have covered in class. Additionally, give examples of 3 different evaluation metrics we have used for classification tasks.

Logistic Regression, Decision Trees, k-NN are some examples

Accuracy, Precision/Recall, TPR/FPR are some examples

2. For ridge regression, describe how choosing a large regularization penalty  $\lambda$  will affect the following. Assume that, before adding the regularizer, we are using a model that is too complex.

- Training error

Training error will be high since model is penalized too much

- Test error

Test error will be high for same reason

- Magnitude of Coefficients

Coefficients will have small magnitude

- Number of 0 coefficients

Ridge does not favor 0 coefficients, so we wouldn't expect any to be 0.

3. For ridge regression, describe how choosing a small regularization penalty  $\lambda$  will affect the following. Assume that, before adding the regularizer, we are using a model that is too complex.

- Training error

Training error will be low because model can overfit to data

- Test error

Test error will be high because model is overfit to data

- Magnitude of Coefficients

Coefficients will be large due to no penalty

- Number of 0 coefficients

Coefficients aren't penalized for being large, none will be 0 most likely.

4. Describe one benefit and one drawback of using k-means++ over k-means.

Benefit: k-means++ generally finds better clusterings

Drawback: Slower initialization

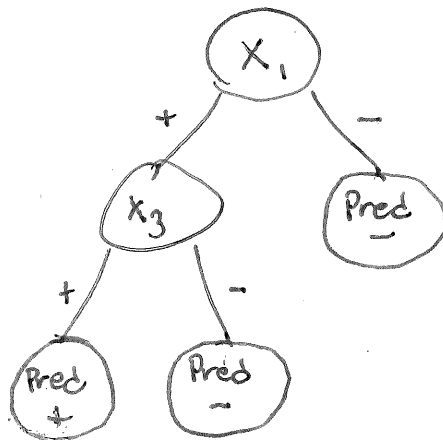
5. What does PCA try to minimize in order to reduce the dimensionality of the dataset.

Reconstruction error

### Short Work

1. For the following dataset, construct a decision tree that classifies the following data. You may assume that there is no depth limit. If at any point there is a tie, you should prefer the feature with the smaller index. Be sure to label all nodes and branches.


$x_1$	$x_2$	$x_3$	$y$
+	-	-	-
+	-	+	+
+	+	+	+
-	-	+	-



2. The following is pseudocode that describes the decision tree algorithm where each feature takes on one of two values. We want to modify this pseudocode to allow for weighted dataset (i.e. boosting). Circle which parts of the code must change and describe what needs to be changed for those parts. Summarize your changes in a sentence or two. There is no need to actually change the code, just write what parts need to change.

```
function DecisionTree(data):  
  if all examples have same label y:  
    return Leaf(y):  
  else:  
    for each feature hi do  
      Data1, Data2 = Split(Data, hi)  
      Error = ClassificationError(Data1)  
              + ClassificationError(Data2)  
    end for  
    h* = choose feature hi that has smallest Error  
    Data1, Data2 = Split(Data, h*)  
    return Branch(h*, DecisionTree(Data1), DecisionTree(Data2))  
  end if  
end function
```

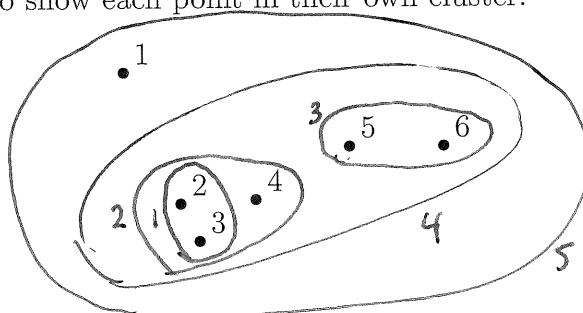
Need to compute  
weighted classification  
error



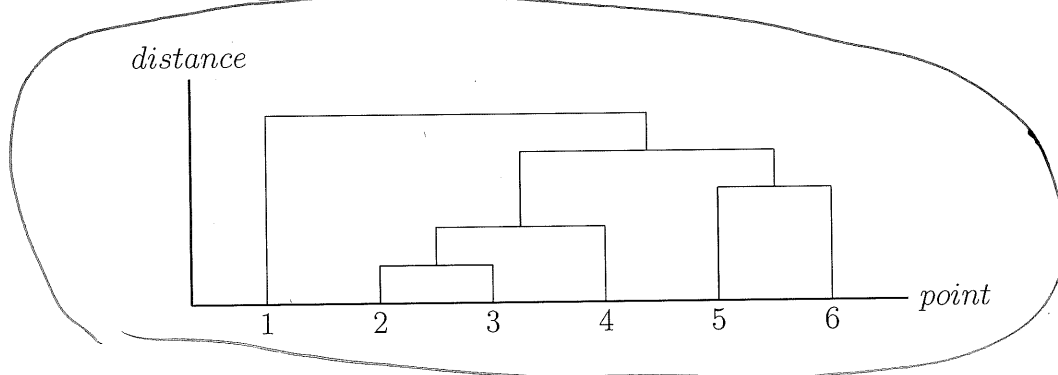
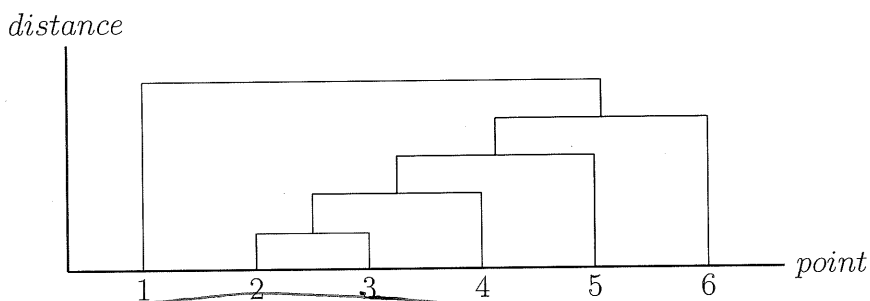
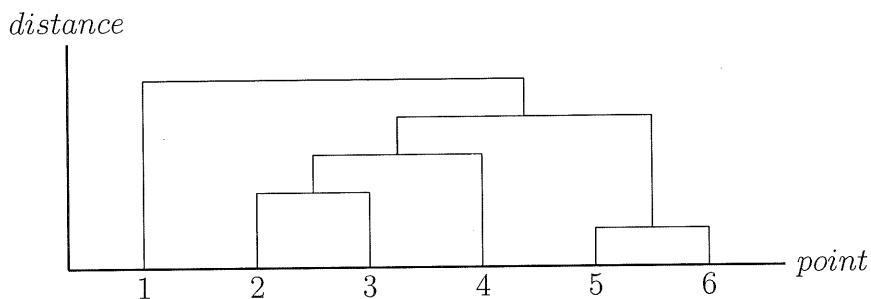
3. Consider the following set of points in  $\mathbb{R}^2$ .

- (a) Draw the order the clusters are formed using agglomerative hierarchical clustering with single linkage. Circle the clusters and label them 1, 2, 3, ... in the order in which they are joined.

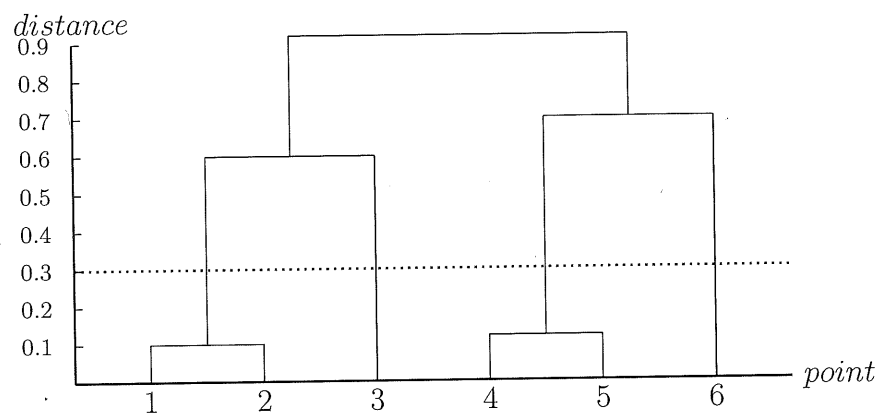
You **do not** need to show each point in their own cluster.



- (b) Which of the following dendrograms matches the *order* in which the clusters are assigned?



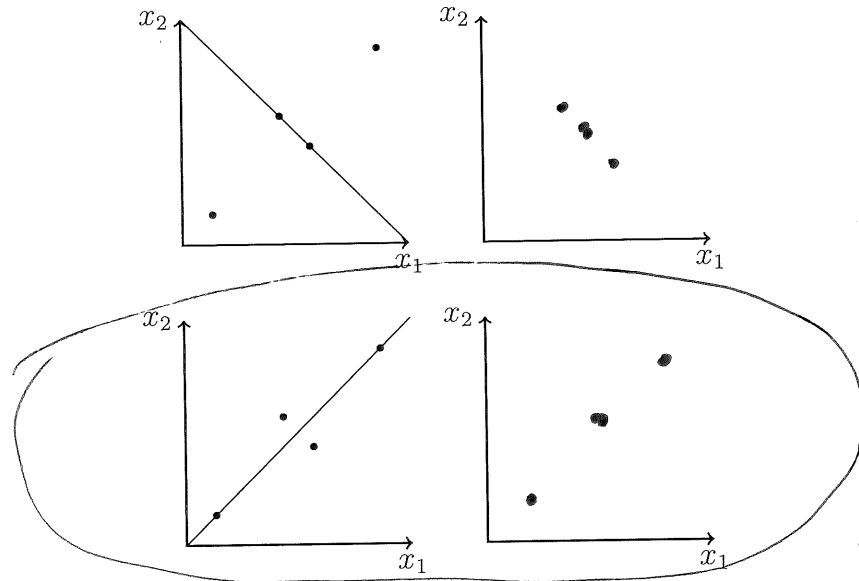
4. Consider the following dendrogram. How many clusters would be reported if were to slice at the dotted line?



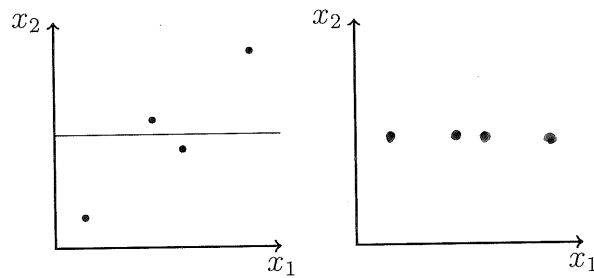
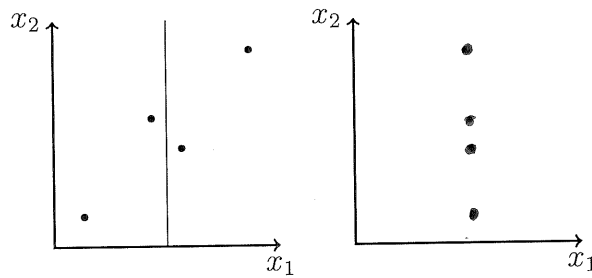
4



5. For the following dataset we show 4 possible candidate lines to project the data onto for PCA. For each line, draw the reconstructed dataset after projecting to that line. Identify which line we would use and state why.



↖ minimizes reconstruction error



6. Suppose we have a movie library consisting of 3 different movies and our system has 4 registered users. Also suppose we have already computed our best estimate for the ratings matrix factorization task and the matrices learned for users and movies are as shown below. Using these matrices, which movie would we recommend to user 2?

$$\hat{R}(U_2, M_1) = 1 \cdot 1 + 0 \cdot 0 = 1$$

$$\hat{R}(U_2, M_2) = 1 \cdot 2 + 0 \cdot 1 = 2$$

$$\hat{R}(U_2, M_3) = 1 \cdot 1 + 0 \cdot 2 = 1$$

1	2
1	0
0	2
1	2

1	2	1
0	1	2

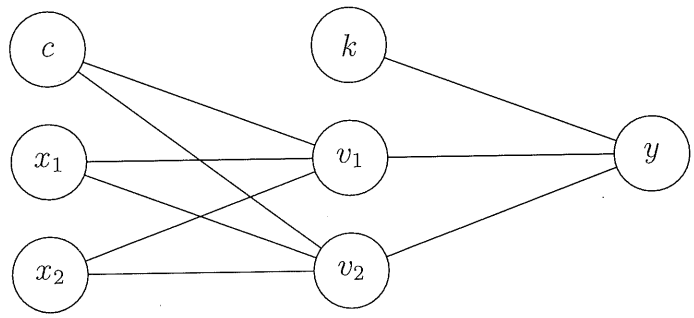
Recommend movie 2

7. Suppose we have the following neural network. What would be the output of the network given the input  $[0, 1]$ ? Assume we are using the ReLU activation function defined by

$$g(x) = \max(0, x)$$

Weights are given by the matrix below:

	$v_1$	$v_2$		$y$
$c$	-0.5	-0.5	$k$	-0.5
$x_1$	1.0	-1.0	$v_1$	1.0
$x_2$	-1.0	1.0	$v_2$	1.0



$$v_1 = g(-0.5 + 1 \cdot x_1 - 1 \cdot x_2) = g(-0.5 + 0 - 1) = g(-1.5) = \max(0, -1.5) = 0$$

$$v_2 = g(-0.5 - 1 \cdot x_1 + 1 \cdot x_2) = g(-0.5 - 0 + 1) = g(0.5) = \max(0, 0.5) = 0.5$$

$$y = g(-0.5 + 1 \cdot v_1 + 1 \cdot v_2) = g(-0.5 + 0 + 0.5) = g(0) = \max(0, 0) = 0$$

8. For the following training dataset, draw the function learned by k-NN regression for the specified values of  $k$ .

