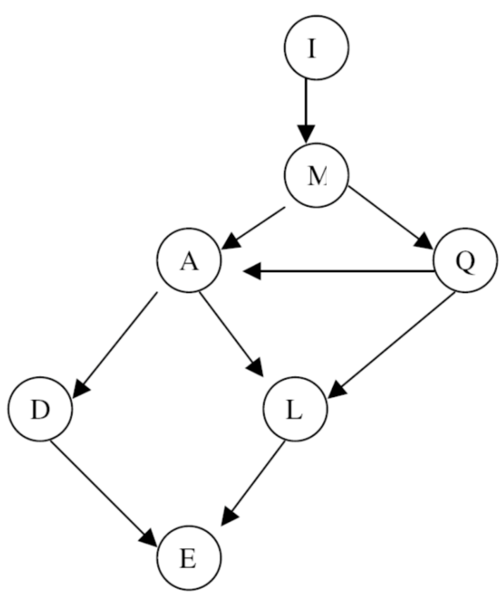
1 Bayesian Networks

1. Draw a Bayesian network among the following binary variables that model the outcome of an election:
   - I : candidate is Incumbent
   - M : has lots of Money for advertising
   - A : uses advertisements that focus on Attacking the candidate's opponent
   - Q : uses advertisements that focus on the candidate's Qualifications
   - L : candidate is Liked
   - D : opponent is Distrusted
   - E : candidate is Elected

   Your network should encode the following beliefs:

   - Incumbents tend to raise lots of money.
   - Money can be used to buy advertising that either focuses on the candidate's qualifications or that attacks the candidate's opponent. But if one does one, there is less money to do the other.
   - Attack advertisements tend to make voters distrust the opponent but they also make the voters tend not to like the candidate.
   - Advertisement focusing on qualifications tends to make the voters like the candidate.
   - Candidates that people like tend to get elected.
   - Candidates whose opponent people distrust tend to get elected



2. For each of the following, say whether it is or is not asserted by the network structure you drew (without assuming anything about the numerical entries in the CPTs).

   1. $P(L|A,Q,D) = P(L|A, Q)$

      Asserted

   2. $P(A|M,Q) = P(A|M)$

      Not asserted

   3. $P(L,D|A,Q) = P(L|A,Q)\ P(D|A,Q)$

      Asserted

2 More Bayesian Networks

Show a Bayesian network structure that encodes the following relationships:

- A is independent of B
- A is dependent on B given C
- A is dependent on D
- A is independent of D given C

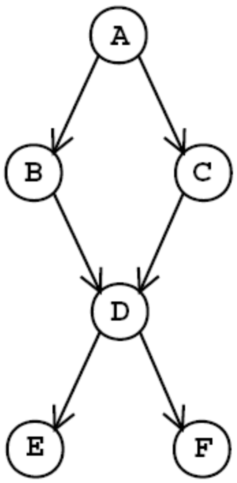
Nodes A and B have no parents

Node C has two parents: A and B

Node D has one parent: C

3 Even More Bayesian Networks

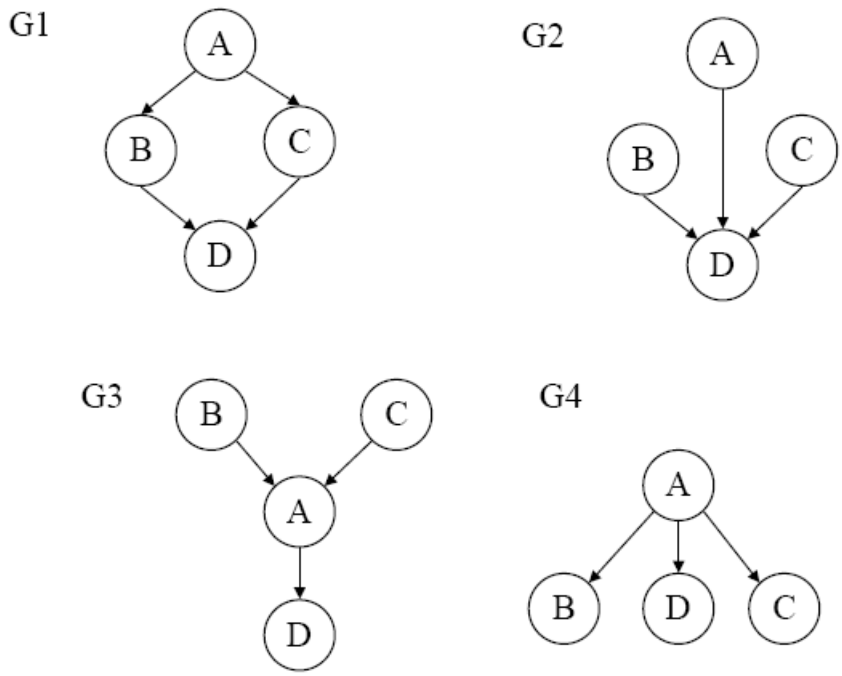
Consider this network:



Which of the following conditional independence assumptions are true?

1. A and E are independent
2. A and E are independent given D - true
3. B and C are independent
4. B and C are independent given A - true
5. B and C are independent given D
6. A and E are independent given B
7. A and E are independent given F
8. B and C are independent given E

4 One Last Question on Bayesian Networks



The following is a list of conditional independence statements. For each statement, name all of the graph structures, G1-G4, or "none" that imply it.

a. A is conditionally independent of B given C – G2

b. A is conditionally independent of B given D - none

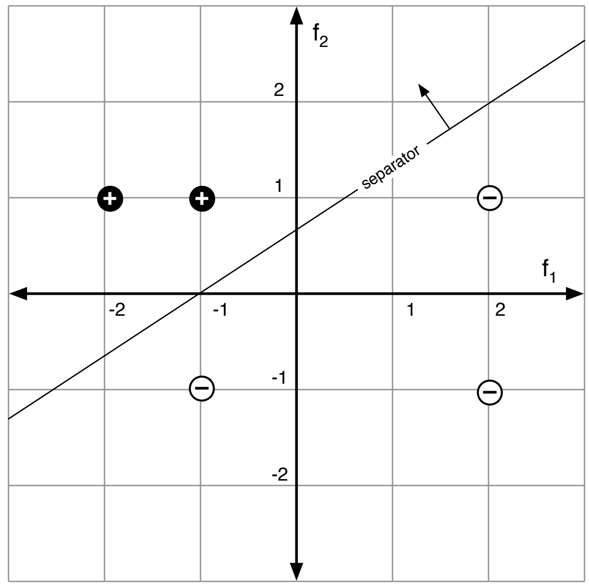
c. B is conditionally independent of D given A – G3, G4

d. B is conditionally independent of D given C - none

e. B is independent of C – G2, G3

f. B is conditionally independent of C given A – G1, G2, G4

3 Maximal Margin Linear Separator (20 points)



Data points are: Negative: (-1, -1) (2, 1) (2, -1) Positive: (-2, 1) (-1, 1)

1. Give the equation of a linear separator that has the maximal geometric margin for the data above. Hint: Look at this geometrically, don't try to derive it formally.

   (a) $w = [-2, 3]$

   (b) $b = -2$

2. Draw your separator on the graph above.

3. What is the value of the smallest geometric margin for any of the points?

   It's the margin for the support vectors divided by the magnitude of w*, that is,* $3/\sqrt{13}$

4. Which are the support vectors for this separator? Mark them on the graph above.

   The support vectors are (-1, 1), (-1, -1) and (2, 1)

6 Naive Bayes (15 points)

Consider a Naive Bayes problem with three features, $x_1 \dots x_3$. Imagine that we have seen a total of 12 training examples, 6 positive (with $y = 1$) and 6 negative (with $y = 0$). Here are the actual points:

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

Here is a table with the summary counts:

	$y = 0$	$y = 1$
$x_1 = 1$	3	3
$x_2 = 1$	3	3
$x_3 = 1$	3	3

1. What are the values of the parameters $R_i(1, 0)$ and $R_i(1, 1)$ for each of the features i (using the Laplace correction)?

   All the parameters are $(3 + 1)/(6 + 2) = 0.5$

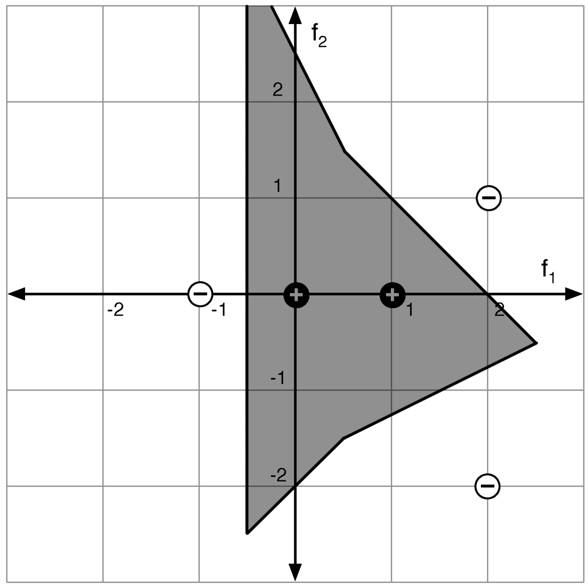
2. If you see the data point $1, 1, 1$ and use the parameters you found above, what output would Naive Bayes predict? Explain how you got the result.

   The prediction is arbitrary since S(0)=S(1) = 1/8

3. Naive Bayes doesn't work very well on this data, explain why.

   The basic assumption of NB is that the features are independent, give the class. In this data set, features 1 and 3 are definitely not independent; the values of these features are opposite for class 0 and equal for class 1. All the information is in these correlations, each feature independently says nothing about the class, so NB is not really applicable. Note that a decision tree would not have any problem with this data set.

2 Nearest Neighbors (8 pts)



Data points are: Negative: (-1, 0) (2, 1) (2, -2) Positive: (0, 0) (1, 0)

1. Draw the decision boundaries for 1-Nearest Neighbors on the graph above. Try to get the integer-valued coordinate points in the diagram on the correct side of the boundary lines.

2. What class does 1-NN predict for the new point: (1, -1.01) Explain why.

   Positive (+) since this is the class of the closest data point (1,0).

3. What class does 3-NN predict for the new point: (1, -1.01) Explain why.

   Positive (+) since it is the majority class of the three closest data points (0,0), (1,0) and (2,-2).

5 Naive Bayes (8 pts)

Consider a Naive Bayes problem with three features, $x_1 \ldots x_3$. Imagine that we have seen a total of 12 training examples, 6 positive (with $y = 1$) and 6 negative (with $y = 0$). Here is a table with some of the counts:

	$y = 0$	$y = 1$
$x_1 = 1$	6	6
$x_2 = 1$	0	0
$x_3 = 1$	2	4

1. Supply the following estimated probabilities. Use the Laplacian correction.

   - $\Pr(x_1 = 1|y = 0) = \frac{6+1}{6+2} = \frac{7}{8}$

   - $\Pr(x_2 = 1|y = 1) = \frac{0+1}{6+2} = \frac{1}{8}$

   - $\Pr(x_3 = 0|y = 0) = 1 - \frac{2+1}{6+2} = \frac{5}{8}$

2. Which feature plays the largest role in deciding the class of a new instance? Why?

   x_3, **because it has the biggest difference in the likelihood of being true for the two different classes. The other two features carry no information about the class.**

6 Learning algorithms (16 pts)

For each of the learning situations below, say what learning algorithm would be best to use, and why.

1. You have about 1 million training examples in a 6-dimensional feature space. You only expect to be asked to classify 100 test examples.

   Nearest Neighbors is a good choice. The dimensionality is low and so appropriate for KNN. For KNN, training is very fast and since there are few classifications, the fact that this will be slow does not matter. With 1 million training examples, neural net and SVM will be extremely expensive to train. Naive Bayes is plausible on computational grounds but likely to be less accurate than KNN.

2. You are going to develop a classifier to recommend which children should be assigned to special education classes in kindergarten. The classifier has to be justified to the board of education before it is implemented.

   A Decision Tree is a good choice since the resulting classifier will need to be understandable to humans.

3. You are working for Am*z*n as it tries to take over the retailing world. You are trying to predict whether customer X will like a particular book, as a function of the input which is a vector of 1 million bits specifying whether each of Am*z*n's other customers liked the book. You will train a classifier on a very large data set of books, where the inputs are everyone else's preferences for that book, and the output is customer X's preference for that book. The classifier will have to be updated frequently and efficiently as new data comes in.

   Naive Bayes is a good choice since it is fast to train and update. The dimensionality is high for Nearest Neighbors and Decision Trees. SVM's have to be re-trained from scratch if the data changes. Neural Nets could be trained incrementally but it will generally take a lot of iterations to change the current settings of the weights.

4. You are trying to predict the average rainfall in California as a function of the measured currents and tides in the Pacific ocean in the previous six months.

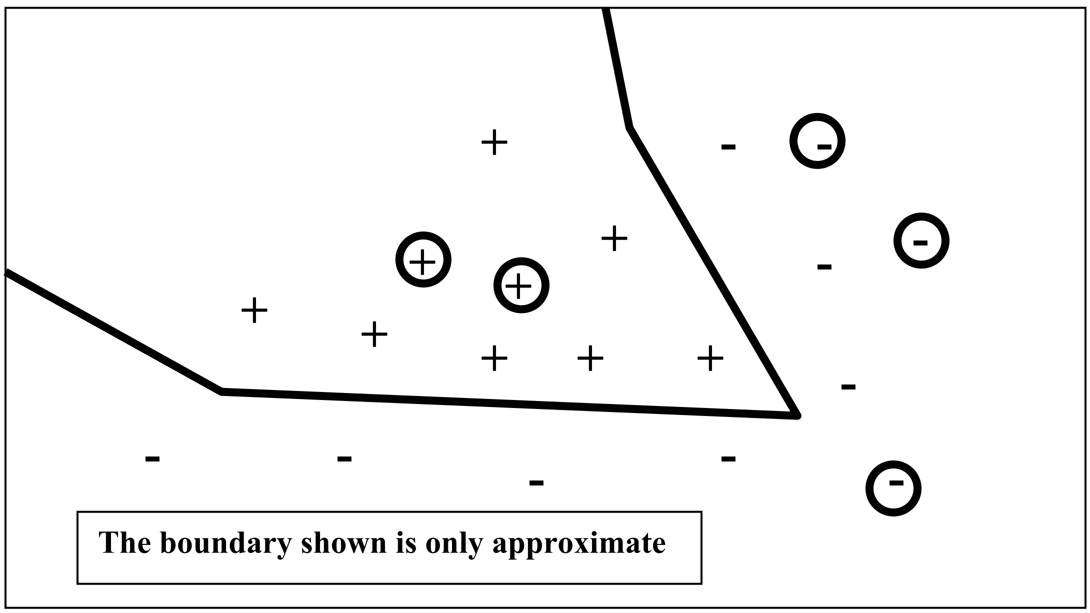
   This is a regression problem; neural nets with linear output functions, regression trees or locally weighted nearest neighbors are all appropriate choices.

Problem 4: Learning (25 points)

Part A: (5 Points)

Since the cost of using a nearest neighbor classifier grows with the size of the training set, sometimes one tries to eliminate redundant points from the training set. These are points whose removal does not affect the behavior of the classifier for any possible new point.

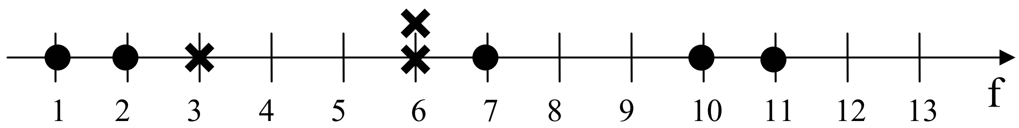
1. In the figure below, sketch the decision boundary for a 1-nearest-neighbor rule and circle the redundant points.



2. What is the general condition(s) required for a point to be declared redundant for a 1-nearest-neighor rule? Assume we have only two classes (+, -). Restating the definition of redundant ("removing it does not change anything") is not an acceptable answer. Hint – think about the neighborhood of redundant points.

Let the Voronoi cell for a training point be the set of points that are closest to that point (as opposed to some other training point). The Voronoi cell of a redundant point touches only on other Voronoi cells of points of the same class.

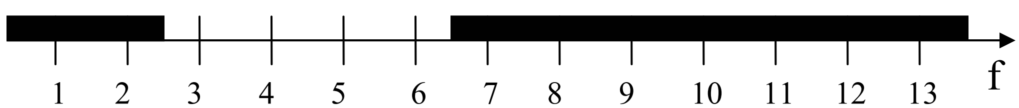
Problem 1: Classification (40 points)



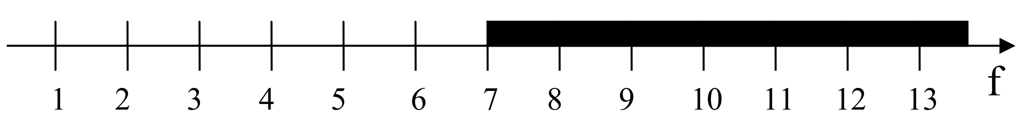
The picture above shows a data set with 8 data points, each with only one feature value, labeled f. Note that there are two data points with the same feature value of 6. These are shown as two X's one above the other, but they really should have been drawn as two X's on top of each other, since they have the same feature value.

Part A: (10 Points)

1. Consider using 1-Nearest Neighbors to classify unseen data points. On the line below, darken the segments of the line where the 1-NN rule would predict an O given the training data shown in the figure above.



2. Consider using 5-Nearest Neighbors to classify unseen data points. On the line below, darken the segments of the line where the 5-NN rule would predict an O given the training data shown in the figure above.



3. If we do 8-fold cross-validation using 1-NN on this data set, what would be the predicted performance? Settle ties by choosing the point on the left. Show how you arrived at your answer.

The point at 1 would be correct, nearest neighbor is at 2
The point at 2 would be correct, nearest neighbor is at 1 (tie)
The point at 3 would be incorrect, nearest neighbor is 2
Both points at 6 would be correct, nearest neighbor is the other point at 6.
The point at 7 would be incorrect, nearest neighbor is 6 (tie)
The point at 10 would be correct, nearest neighbor is 11
The point at 11 would be correct, nearest neighbor is 10

So,6 correct, 2 incorrect => 75% would be predicted performance

Problem 2: Overfitting (20 points)

For each of the supervised learning methods that we have studied, indicate how the method could overfit the training data (consider both your design choices as well as the training) and what you can do to minimize this possibility. There may be more than one mechanism for overfitting, make sure that you identify them all.

Part A: Nearest Neighbors (5 Points)

1. How does it overfit?

   Every point in dataset (including noise) defines its own decision boundary.

   The distance function can be chosen to do well on training set but less well on new data.

2. How can you reduce overfitting?

   Use k-NN for larger k

   Use cross-validation to choose k and the distance function

Part B: Decision Trees (5 Points)

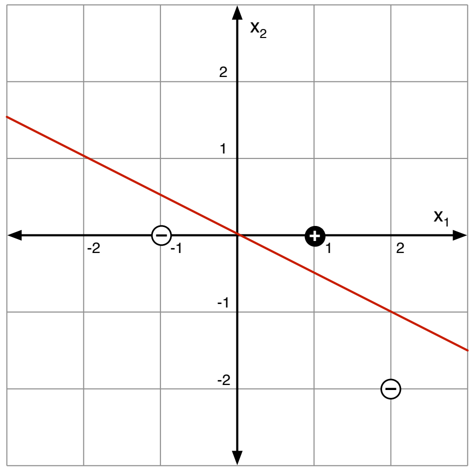
1. How does it overfit?

   By adding new tests to the tree to correctly classify every data point in the training set.

2. How can you reduce overfitting?

   By pruning the resulting tree based on performance on a validation set.

3 Perceptron (7 pts)



Data points are: Negative: (-1, 0) (2, -2) Positive: (1, 0). Assume that the points are examined in the order given here. Recall that the perceptron algorithm uses the extended form of the data points in which a 1 is added as the 0th component.

1. The linear separator obtained by the standard perceptron algorithm (using a step size of 1.0 and a zero initial weight vector) is (0 1 2). Explain how this result was obtained.

   The perceptron algorithm cycles through the augmented points, updating weights according to the update rule $w_{\text{new}} = w + y \cdot x$ after misclassifying points. The intermediate weights are given in the table below.

Test point	misclassified?	Updated weights
Initial weights		**0 0 0**
-: (1 -1 0)	**yes**	**-1 1 0**
-: (1 2 -2)	**yes**	**-2 -1 2**
+: (1 1 0)	**yes**	**-1 0 2**
-: (1 -1 0)	**no**	
-: (1 2 -2)	**no**	
+: (1 1 0)	**yes**	**0 1 2**
-: (1 -1 0)	**no**	
-: (1 2 -2)	**no**	
+: (1 1 0)	**no**	

2. What class does this linear classifier predict for the new point: (2.0, -1.01)

   The margin of the point is -0.01, so it would be classified as negative.

3. Imagine we apply the perceptron learning algorithm to the 5 point data set we used on Problem 1: Negative: (-1, 0) (2, 1) (2, -2), Positive: (0, 0) (1, 0). Describe qualitatively what the result would be.

   The perceptron algorithm would not converge since the 5 point data set is not linearly separable.

6 Perceptron (8 points)

The following table shows a data set and the number of times each point is misclassified during a run of the perceptron algorithm, starting with zero weights. What is the equation of the separating line found by the algorithm, as a function of x_1, x_2, and x_3? Assume that the learning rate is 1 and the initial weights are all zero.

x_1	x_2	x_3	y	times misclassified
2	3	1	+1	12
2	4	0	+1	0
3	1	1	-1	3
1	1	0	-1	6
1	2	1	-1	11

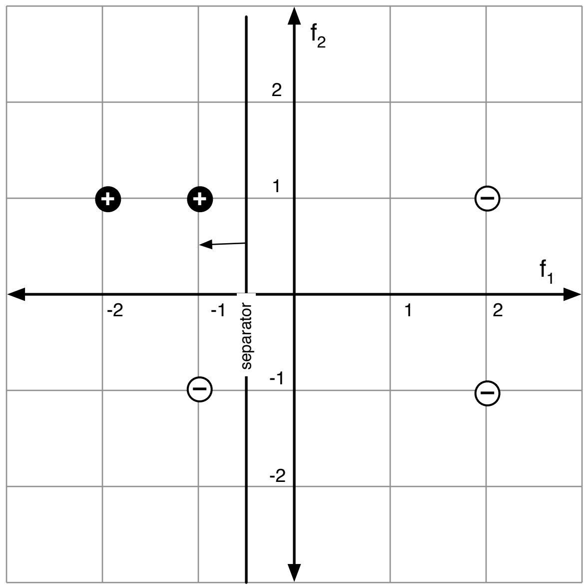
$$\begin{aligned}\bar{w} &= \eta \sum_{i=1}^{m} \alpha_i y^i \bar{x}^i \\ &= (12)(1)(1,2,3,1) + (3)(-1)(1,3,1,1) + (6)(-1)(1,1,1,0) + (11)(-1)(1,1,2,1) \\ &= (-8,-2,5,-2)\end{aligned}$$

So the equation of the separating line is

$$-2x_1 + 5x_2 - 2x_3 - 8 = 0$$

6.034 Quiz 4 Solutions, Spring 2006

1 Perceptron (20 points)



Data points are: Negative: (-1, -1) (2, 1) (2, -1) Positive: (-2, 1) (-1, 1)

Recall that the perceptron algorithm uses the extended form of the data points in which a 1 is added as the 0th component.

1. Assume that the initial value of the weight vector for the perceptron is $[0, 0, 1]$, that the data points are examined in the order given above and that the rate (step size) is 1.0. Give the weight vector after one iteration of the algorithm (one pass through all the data points):

   Only point $x_2 = (2, 1)$ is misclassified. Using the extended form $x_2' = (1, 2, 1)$, we have

   $$[0, 0, 1] * x_2' = +1$$

   We update the weight vector using the extended form (times $y_2 = -1$):

   $$w \leftarrow w + y_2 x_2'$$

   and we get the new weight vector

   $$w = [-1, -2, 0]$$

2. Draw the separator corresponding to the weights after this iteration on the graph at the top of the page.

3. Would the algorithm stop after this iteration or keep going? Explain.

   No. The new weight vector misclassifies the negative point (-1, -1), whose margin will be +1.

4. If we add a positive point at (1,-1) to the other points and retrain the perceptron, what would the perceptron algorithm do? Explain.

   The data is no longer linearly separable and so the perceptron would loop forever.