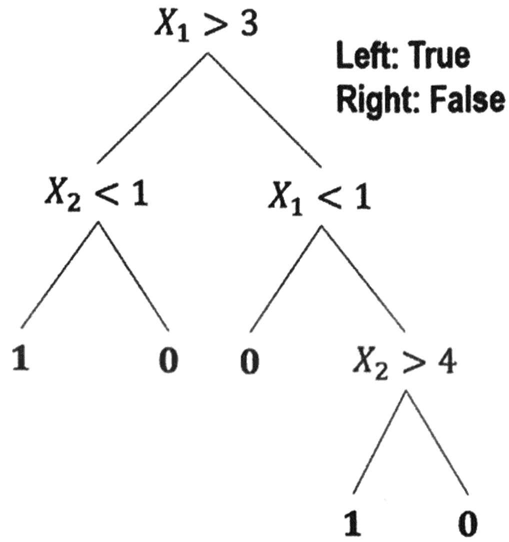
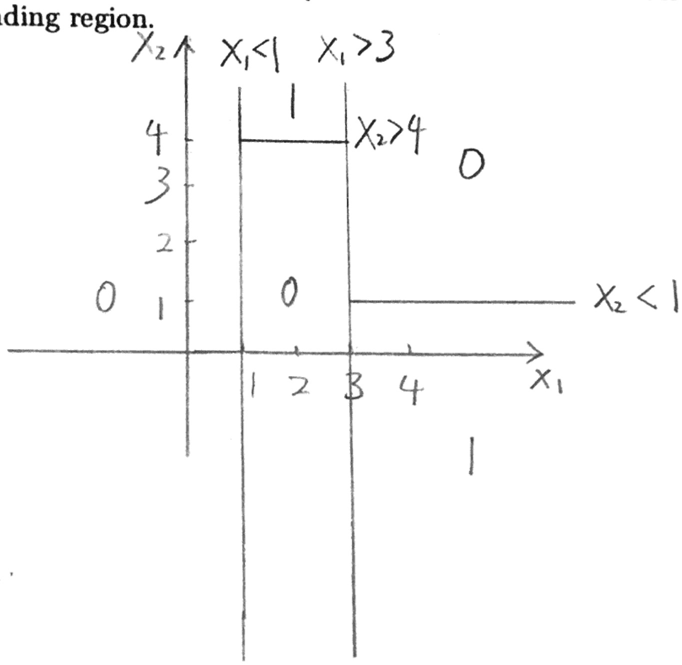
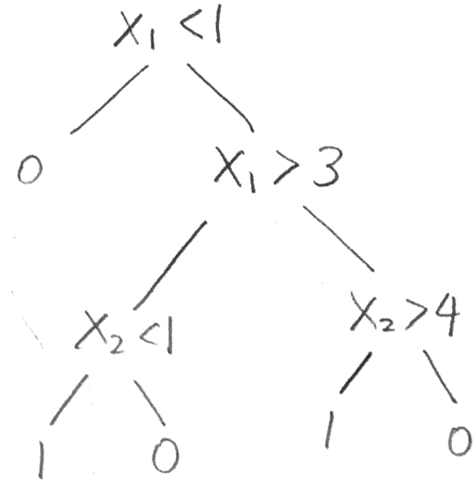
1. (Decision Trees, 20 points) Consider the following decision tree:



a. Draw the decision boundaries defined by this tree. Each leaf is labeled with a number. Write this number in the corresponding region.



b. Give another tree that is syntactically different but defines the same decision boundaries.



c. When a decision tree is grown to full depth, it is more likely to fit the noise in the data. True or False? Explain.

True.

Decision tree can fit every data point as a leaf.

2. (Naïve Bayes, 15 points) Consider the following data set, where the classification task is to predict if a car is going to be bought.

Color	Type	Origin	Buy
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

a. Assume the Naïve Bayes condition and estimate the parameters using maximum likelihood estimation.

Color = {Red, Yellow}, Type = {Sports, SUV},

Origin {Domestic, Imported}, Buy = {Yes, No}

$P(B=Y) = \frac{5}{10}$

$P(B=N) = \frac{5}{10}$

$P(C=R|B=Y) = \frac{3}{5}$ $P(C=R|B=N) = \frac{2}{5}$

$P(C=Y|B=Y) = \frac{2}{5}$ $P(C=Y|B=N) = \frac{3}{5}$

$P(T=Sp|B=Y) = \frac{4}{5}$ $P(T=Sp|B=N) = \frac{2}{5}$

$P(T=S|B=Y) = \frac{1}{5}$ $P(T=S|B=N) = \frac{3}{5}$

$P(O=D|B=Y) = \frac{2}{5}$ $P(O=D|B=N) = \frac{3}{5}$

$P(O=I|B=Y) = \frac{3}{5}$ $P(O=I|B=N) = \frac{2}{5}$

b. For the same domain, apply Laplace correction and estimate the parameters.

$P(B=Y) = \frac{5+1}{10+2}$

$P(B=N) = \frac{5+1}{10+2}$

$P(C=R|B=Y) = \frac{3+1}{5+2}$ $P(C=R|B=N) = \frac{2+1}{5+2}$

$P(C=Y|B=Y) = \frac{2+1}{5+2}$ $P(C=Y|B=N) = \frac{3+1}{5+2}$

$P(T=Sp|B=Y) = \frac{4+1}{5+2}$ $P(T=S|B=N) = \frac{2+1}{5+2}$

$P(T=S|B=Y) = \frac{1+1}{5+2}$ $P(T=S|B=N) = \frac{3+1}{5+2}$

$P(O=D|B=Y) = \frac{2+1}{5+2}$ $P(O=D|B=N) = \frac{3+1}{5+2}$

$P(O=I|B=Y) = \frac{3+1}{5+2}$ $P(O=I|B=N) = \frac{2+1}{5+2}$

c. The logistic function is given by

$$\sigma(x) = \frac{1}{1+e^{-x}}.$$

Show that $\frac{d}{dx}\sigma(x) = \sigma(x) \cdot (1 - \sigma(x))$.

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$$= -\frac{1}{(1+e^{-x})^2} \cdot \frac{d}{dx}(1+e^{-x})$$

$$= -\frac{1}{(1+e^{-x})^2} \cdot e^{-x} \cdot (-1)$$

$$= \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$

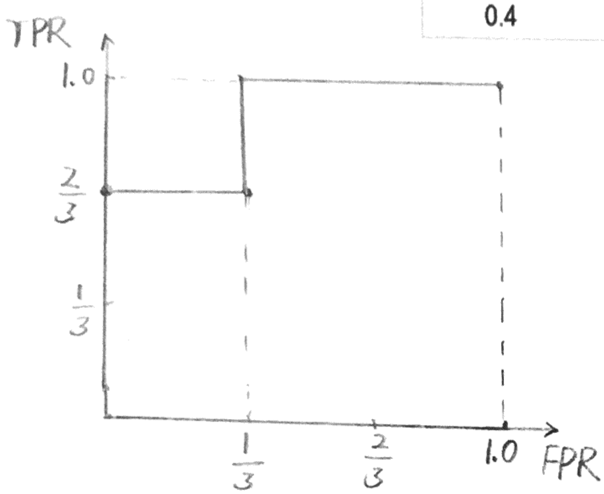
$$= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= \sigma(x) \cdot (1 - \sigma(x))$$

3. **(Evaluation, 15 points)**

a. Let the following predictions be the output of a probabilistic classifier. Draw the ROC curve for this prediction task.

Predicted Probability	Correct Class
0.96	+
0.80	+
0.76	-
0.66	+
0.52	-
0.4	-



b. What is the AUC-ROC for this classifier?

$\frac{8}{9}$

4. (Short Questions, 30 points) For the following questions, select your answer and **explain your reasoning. You will get no points without an explanation.**

a. Suppose you are given a data set 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. **True** or **False**?

True. Good performance with pruning.

b. 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. **True** or **False**?

False. Recall is a better measurement of the performance.

c. 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. **True** or **False**?

False. The ability of generalization is important

d. In logistic regression, we model the odds ratio $(\frac{p}{1-p})$ as a linear function. **True** or **False**?

False.

$$\log\left(\frac{p}{1-p}\right) = wx$$

$$\frac{\frac{1}{1+e^{-wx}}}{1-\frac{1}{1+e^{-wx}}} = e^{wx}$$

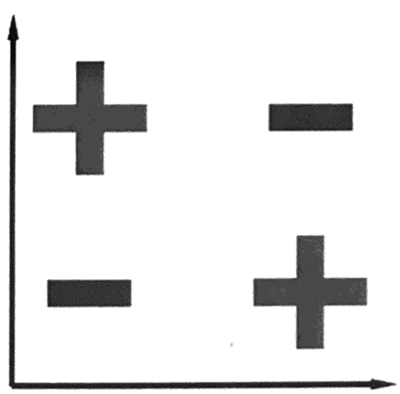
e. If you train a linear regression estimator with only half the data, its bias is smaller. **True** or **False**?

False. Bias is smaller with larger data set in general.

f. Because decision trees learn to classify discrete-valued outputs instead of real-valued functions, it is impossible for them to overfit. **True** or **False**?

False. It can fit every data point as a leaf.

g. Which of the following classifiers can perfectly classify the following data set? (i) support vector machines, (ii) logistic regression, (iii) perceptron, (iv) decision trees.



(i) with Gaussian kernel
(iv)

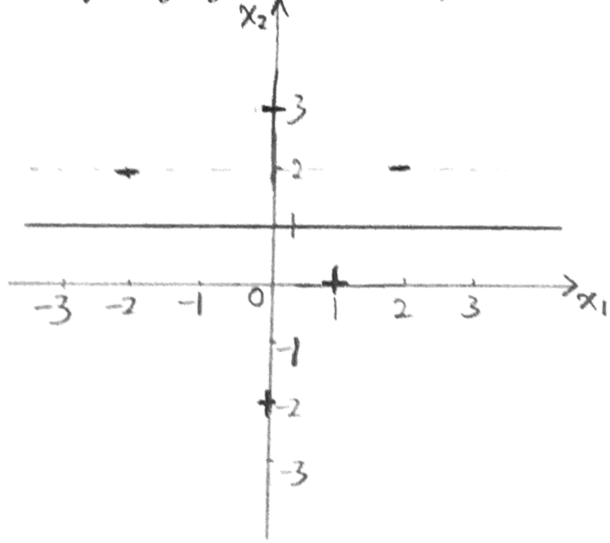
h. When the hypothesis space is richer, over fitting is more likely. **True** or **False**?

True. Variance is higher, generalization may be worse.

5. **(Support Vector Machines, 20 points)** Consider the data set below, with features x_1, x_2 and labels y:

x_1	x_2	y
1	0	+1
-2	2	−1
0	3	−1
0	-2	+1
2	2	−1

a. Find the linear SVM classifier for the data set given above in the form $w_1x_1 + w_2x_2 + b = 0$. (*Hint: Consider plotting the data set and finding a geometric solution*).



$w_1 = 0$

$w_2 = -1$

$b = 1$

b. What is the margin of the linear SVM classifier from the previous question? What are the support vectors for this classifier?

margin : $x_2 = 2$, $x_2 = 0$

support vectors: $(-2, 2)$, $(2, 2)$, $(1, 0)$

c. One of the most commonly used kernels is the Gaussian kernel

$$\kappa(\mathbf{x}, \mathbf{z}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{z}\|^2\right).$$

Consider three points $\mathbf{x}$, $\mathbf{z}_1$ and $\mathbf{z}_2$. Geometrically, we know that $\mathbf{x}$ and $\mathbf{z}_1$ are **very far from each other**, and that $\mathbf{x}$ and $\mathbf{z}_2$ are **very close to each other**. Which one of the following is true, and why?

(a) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to 1 and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to 0.

(b) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to 0 and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to 1.

(c) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to c_1, $c_1 \gg 1$ and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to c_2, $c_2 \ll 0$, where $c_1, c_2 \in \mathbb{R}$.

(d) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to c_1, $c_1 \ll 0$ and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to c_2, $c_2 \gg 1$, where $c_1, c_2 \in \mathbb{R}$.

(b) is true

z_1, x are far $\Rightarrow \|z_1 - x\|^2$ is large $(+\infty)$ $\Rightarrow K(z_1, x) = \exp(-\gamma\|z_1 - x\|^2)$ is close to 0

z_2, x are close $\Rightarrow \|z_2 - x\|^2$ is small (0) $\Rightarrow K(z_2, x) = \exp(-\gamma\|z_2 - x\|^2)$ is close to 1.