一、Short Questions

1. Does a 2-class Gaussian Naive Bayes classifier with parameter $\mu_{1k}, \sigma_{1k}, \mu_{2k}, \sigma_{2k}$ for attributes $k = 1, \ldots, m$ have exactly the same representational power as logistic regression (i.e., a linear decision boundary), given no assumptions about the variance values σ_{ik}?

2. For linear separable data, can a small slack penalty ("C") hurt the training accuracy when using a linear SVM (no kernel)? If so, explain how. If not, why not?

3. PCA and spectral clustering perform eigen-decomposition on two different matrices. However, the size of these two matrices are the same.

4. The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

5. Since classification is a special case of regression, logistic regression is a special case of linear regression.

二、

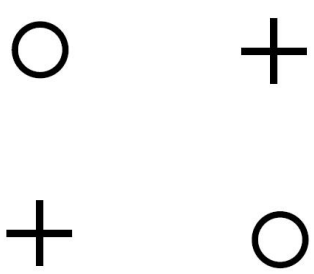
1. Given 3 data points in 2-d space, (1,1), (2,2) and (3,3),

(a) What is the first principle component?

(b) If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

(c) For the projected data in (b), now if we represent them in the original 2-d space, what is the reconstruction error?

2. Consider the following data set:



Circle all of the classifiers that will achieve zero training error on this data set. (You may circle more than one)

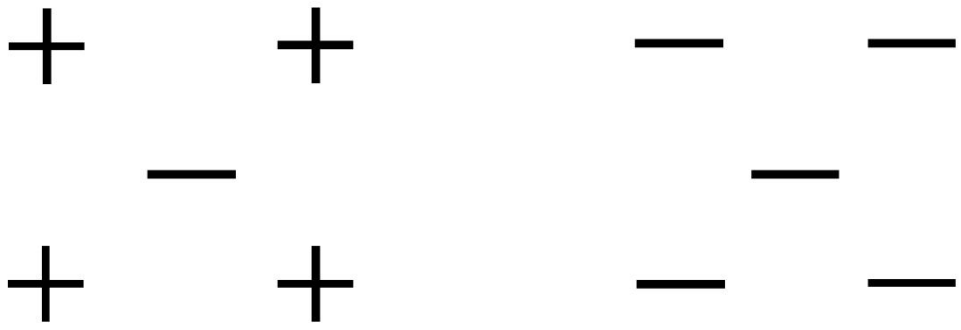
(a) Logistic regression

(b) SVM (quadratic kernel)

(c) Decision trees

(d) 3-NN classifier

3. For the following dataset, circle the classifier which has larger Leave-One-Out Cross-validation error.



(a) 1-NN

(b) 3-NN

4. Construct a one dimensional classification dataset for which the Leave-one-out cross validation error of the 1-Nearest Neighbors algorithm is always 1. Stated another way, the 1-NN algorithm never correctly predicts the held out point.

5. Given *n* linearly independent feature vectors in *n* dimensions, show that for any assignment to the binary labels, you can always construct a linear classifier with weight vector *w* which separates the points. Assume that the classifier has the form sign(*wx*). Note that a square matrix composed of linearly independent rows is invertible.

6. Let $F(x) = \omega_0 + \sum_{j=1}^{d} \omega_j x_j$ and $L(yF(x)) = \frac{1}{1+\exp(yF(x))}$. Suppose you use gradient descent to obtain the optimal parameters ω_0 and ω_j. Give the update rules for these parameters.

7. Is it possible to construct a Bayes classifier for one input x such that when it is used, it will predict

- Class 1 if $x < -1$
- Class 2 if $-1 < x < 1$
- Class 1 if $1 < x$?

If so, how?

三、Bayes Rule

(a) I give you the following fact:

$$P(A \mid B) = 2/3$$

Do you have enough information to compute $P(B \mid A)$? If not, write "not enough info". If so, compute the value of $P(B \mid A)$.

(b) Instead, I give you the following facts:

$$P(A \mid B) = 2/3$$

$$P(A \mid \sim B) = 1/3$$

Do you have enough information to compute $P(B \mid A)$? If not, write "not enough info". If so, compute the value of $P(B \mid A)$.

(c) Instead, I give you the following facts:

$$P(A \mid B) = 2/3$$

$$P(A \mid \sim B) = 1/3$$

$$P(B) = 1/3$$

Do you have enough information to compute $P(B \mid A)$? If not, write "not enough info". If so, compute the value of $P(B \mid A)$.

(d) Instead, I give you the following facts:

$$P(A \mid B) = 2/3$$

$$P(A \mid \sim B) = 1/3$$

$$P(B) = 1/3$$

$$P(A) = 4/9$$

Do you have enough information to compute $P(B \mid A)$? If not, write "not

enough info". If so, compute the value of $P(B \mid A)$.

四、Decision Trees

We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or low) and whether or not they studied.

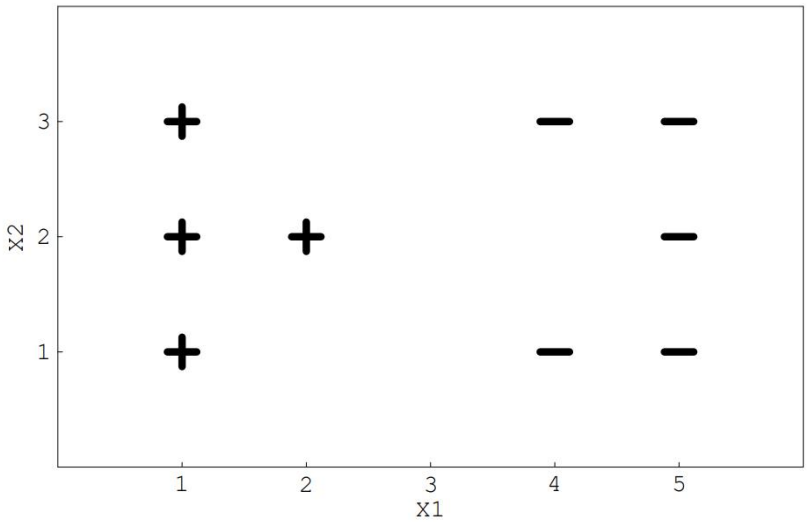
GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

For this problem, you can write your answers using $\log_2$, but it may be helpful to note that $\log_2 3 \approx 1.6$.

1. What is the entropy H(Passed)?
2. What is the entropy H(Passed|GPA)?
3. What is the entropy H(Passed|Studied)?
4. Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.

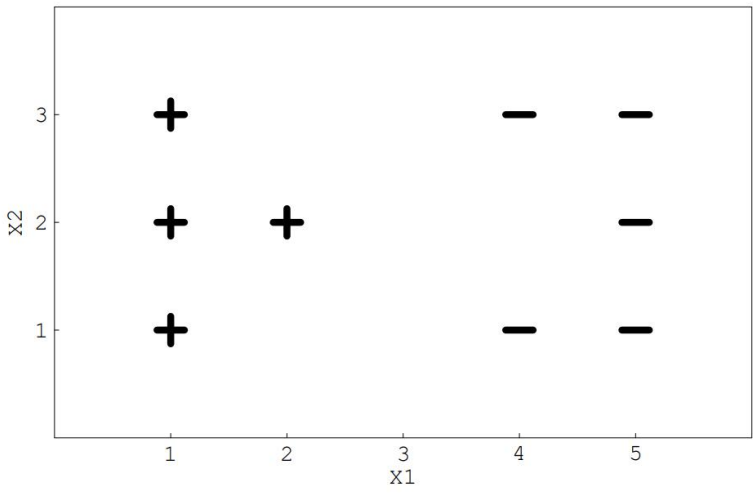
五、Support Vector Machine

1. Suppose that we are using a linear SVM (i.e., no kernel), with some large C value, and are given the following data set.



Draw the decision boundary of linear SVM. Give a brief explanation.

2. In the following image, circle the points such that removing that example from the training set and retraining SVM, we would get a different decision boundary than training on the full sample.

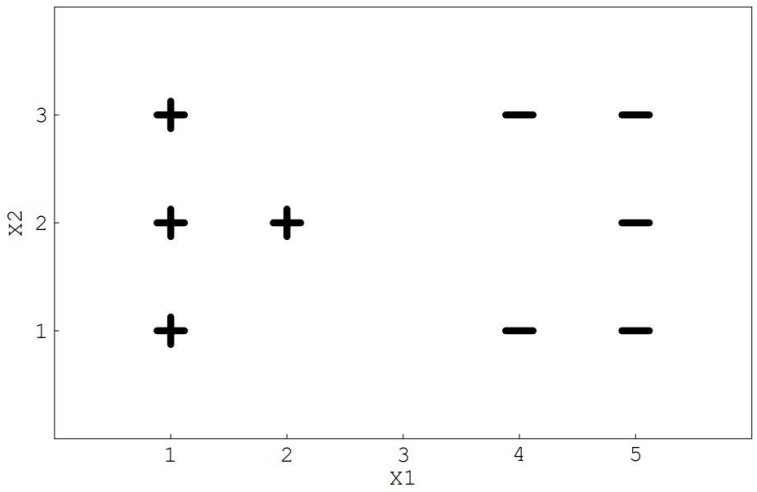


You do not need to provide a formal proof, but give a one or two sentence explanation.

3. Suppose instead of SVM, we use regularized logistic regression to learn the classifier. That is,

$$(w,b) = \arg\min_{w\in\mathbb{R}^2, b\in\mathbb{R}} \frac{\|w\|^2}{2} - \sum_i \mathbb{1}[y^{(i)}=0] \ln \frac{1}{1+e^{(w\cdot x^{(i)}+b)}} + \mathbb{1}[y^{(i)}=1] \ln \frac{e^{(w\cdot x^{(i)}+b)}}{1+e^{(w\cdot x^{(i)}+b)}}.$$

In the following image, circle the points such that removing that example from the training set and running regularized logistic regression, we would get a different decision boundary than training with regularized logistic regression on the full sample set.



4. Suppose we have a kernel $K(\cdot,\cdot)$, such that there is an implicit high-dimensional feature map $\phi: R^d \to R^D$ that satisfies $\forall x, z \in R^d, K(x,z) = \phi(x)\cdot\phi(z)$, where $\phi(x)\cdot\phi(z) = \sum_{i=1}^{D} \phi(x)_i \cdot \phi(z)_i$ is the dot product in the D-dimensional space. Show how to calculate the Euclidean distance in the D-dimensional space

$$\|\phi(x) - \phi(z)\| = \sqrt{\sum_{i=1}^{D} (\phi(x)_i - \phi(z)_i)^2}$$

Without explicitly calculating the values in the D-dimensional vectors.

For this question, you should provide a formal proof.