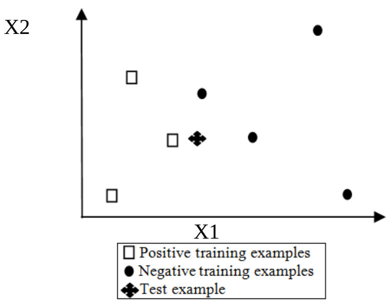
<u>**Midterm 2018/2019**</u>

Question 1: Mark each statement with T or F in the right side: *[5 marks]*

1) In supervised learning, The learning algorithm detects similarity between different training data inputs	(F)
2) We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.	(F)
3) When a decision tree is grown to full depth, it is more likely to fit the noise in the data.	(T)
4) When the feature space is larger, over fitting is more likely.	(T)
5) Since classification is a special case of regression, logistic regression is a special case of linear regression.	(F)
6) The Gradient descent will always find the global optimum	(F)
7) Overfitting Indicates limited generalization	(T)
8) In Support Vector Machines (SVM) ,Inputs are mapped to lower dimensional space where data becomes likely to be linearly separable	(F)
9) When the trained system matches the training set perfectly, overfitting may occur	(T)
10) Algorithms for supervised learning are not directly applicable for unsupervised learning	(T)

Question 2

In Figure we depict training data and a single test point for the task of classification given two continuous attributes X1 and X2. For each value of k, circle the label predicted by the k-nearest neighbor classifier for the depicted test point.



1. Predicted label for k = 1:
   (a) positive (b) negative
2. Predicted label for k = 3:
   (a) positive (b) negative
3. Predicted label for k = 5:
   (a) positive (b) negative

Question 3

Assume the following data

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Construct a parametric classifier using Naïve byes to predict whether this person with a new instance

X= (Given Birth= "Yes", Can Fly= "no", Live in water = "Yes", Have legs="no")

Will be mammals or non-mammals.

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

Question4 with short answer

11) The training error of 1-NN classifier is 0. (true/false) Explain
True: Each point is its own neighbor, so 1-NN classifier achieves perfect classification on training data.

12) Consider a naive Bayes classifier with 3 boolean input variables, $X1;X2$ and $X3$, and one Boolean output, Y . How many parameters must be estimated to train such a naive Bayes classifier? (list them) .

Solutions:
For a naive Bayes classifier, we need to estimate P(Y=1), $P(X1 = 1|y = 0)$*;* $P(X2 = 1|y = 0)$, $P(X3 = 1|y = 0)$, $P(X1 = 1|y = 1)$*;* $P(X2 = 1|y = 1)$*;* $P(X3 = 1|y = 1)$*. Other probabilities can be obtained with the constraint that the probabilities sum up to 1.*

So we need to estimate 7 or 8 parameters.

13) The depth of a learned decision tree can be larger than the number of training examples used to create the tree. . (true/false) Explain

False: Each split of the tree must correspond to at least one training example, therefore, if there are n training examples, a path in the tree can have length at most n

14) We consider the following models of logistic regression for a binary classification with a sigmoid function

$$g(z) = \frac{1}{1+e^{-z}}$$

- Model 1: $P(Y = 1 \mid X, w_1, w_2) = g(w_1 X_1 + w_2 X_2)$
- Model 2: $P(Y = 1 \mid X, w_1, w_2) = g(w_0 + w_1 X_1 + w_2 X_2)$

We have three training examples:

$$\begin{array}{lll} x^{(1)} = [1,1]^T & x^{(2)} = [1,0]^T & x^{(3)} = [0,0]^T \\ y^{(1)} = 1 & y^{(2)} = -1 & y^{(3)} = 1 \end{array}$$

Does it matter how the third example is labeled in Model 1? i.e., would the learned value of w = (w1, w2) be different if we change the label of the third example to -1? Does it matter in Model 2? Briefly explain your answer. (Hint: think of the decision boundary on 2D plane.)

It does not matter in Model 1 because $x^{(3)} = (0, 0)$ makes $w_1x_1 + w_2x_2$ always zero and hence the likelihood of the model does not depend on the value of w. But it does matter in Model 2.

15) Briefly describe the difference between a *maximum likelihood* hypothesis and a *maximum a posteriori* hypothesis.

Solutions:

ML: maximize the data likelihood given the model, i.e., $\arg\max_W P(Data|W)$

MAP: $\arg\max_W P(W|Data)$