Question 1 [8 pts] A grocery store sells apples from 3 different countries (country A, country B and country C). All the apples from country A are sweet, while 75% of the apples from country B are sweet, and 50% of the apples from country C are sweet. Unfortunately, the sign in the grocery store that would normally indicate where the apples are from is missing. Nevertheless, you buy a bag of apples. Before tasting any apple, you believe that they are from country A with probability 0.5, country B with probability 0.25 and country C with probability 0.25. After tasting 5 apples, you noticed that 4 of them are sweet and 1 of them are not sweet.

a) [4 pts] Bayesian learning: what is your posterior belief (after tasting the 5 apples mentioned above) that the apples are from each country? Find the prior, the likelihood, the unnormalized posterior, and the normalized posterior.

prior: Pr(A)=0.5, Pr(B)=0.25, Pr(C)=0.25

likelihood:
Pr(sweet|A)=1, Pr(sweet|B)=0.75, Pr(sweet|C)=0.5

Unnormalized posterior:
Pr(A|4sweets,1~sweet)=0,
Pr(B|4sweets,1~sweet)=$0.25(0.75)^4(0.25)^1$=0.01975,
Pr(C|4sweets,1~sweet)=$0.25(0.5)^5$=0.00781

Normalized posterior:
Pr(A|4sweets,1~sweet)=0/(0.01975+0.00781):0,
Pr(B|4sweets,1~sweet)=0.01975/(0.01975+0.00781)=0.7168:,
Pr(C|4sweets,1~sweet)=0.00781/(0.01975+0.00781)=0.2832
Pr(A|4sweets,1~sweet)+ Pr(B|4sweets,1~sweet)+ Pr(C|4sweets,1~sweet)=1

b) [4 pts] What is the probability that the next apple that you will taste is sweet according to the Bayesian learning prediction?

Pr(sweet|4sweets,1~sweet)= Pr(sweet|A) Pr(A|4sweets,1~sweet)+ Pr(sweet|B) Pr(B|4sweets,1~sweet)+ Pr(sweet|B) Pr(B|4sweets,1~sweet)=0.6792

Question 2 [12 pts] Consider the following dataset. The input space has one dimension and there are two classes (+ and -): (0,+), (1,+), (0.9,+), (0.5,+), (1.5, -), (0.7, -), (1.2, -), (0.95,-)

a) [8 pts] Suppose that we are training a mixture of Gaussians model by maximum likelihood. What are the prior probabilities of each class Pr(+) and Pr(-)? What is the mean of the conditional distribution of each class and what is the variance (assuming that both class conditional distributions have the same variance)?

Pr(+)=4/8=0.5; Pr(-)=0.5

$\mu_+ = \frac{0+1+0.9+0.5}{4} = 0.6, \mu_- = \frac{1.5+0.7+1.2+0.95}{4} = 1.0875$

$S_+ = 0.25[(0-0.6)^2 + (0.9-0.6)^2 + (1-0.6)^2 + (0.5-0.6)^2] = 0.155$

$S_- = 0.25[(1.5 - 1.0875)^2 + (0.7 - 1.0875)^2 + (1.2 - 1.0875)^2 + (0.95 - 1.0875)^2] = 0.08797$
$\Sigma = 0.5(0.155 + 0.08797) = 0.1215$

b) [4 pts] What is the probability that 0.8 belongs to class +?

$$\Pr(+|x) = \frac{\Pr(+)\Pr(x|+)}{\Pr(+)\Pr(x|+) + \Pr(-)\Pr(x|-)}$$
$$\Pr(+) = \Pr(-) = 0.5$$
$$\Pr(x = 0.92|+) = \exp\left(-\frac{0.5(0.8 - 0.6)(0.8 - 0.6)}{0.1215}\right) = 0.8482$$
$$\Pr(x = 0.92|-) = \exp\left(-\frac{0.5(0.8 - 1.0875)(0.8 - 1.0875)}{0.1215}\right) = 0.7112$$
$$\Pr(+|0.92) = \frac{0.5(0.8482)}{0.5(0.8482 + 0.7112)} = 0.5439$$

Question 3 [12 pts] Indicate whether each statement is true or false. No justification required.

a) [2 pts] In linear regression, minimizing the squared loss and maximizing the likelihood of the data under the assumption of Gaussian noise gives the same result. T

b) [2 pts] In k-nearest neighbors, a small k may lead to underfitting while a large k may lead to overfitting. F

c) [2 pts] Linear models with suitable basis functions can be used for non-linear classification and regression. T

d) [2 pts] Logistic regression is a classification technique, not a regression technique. T

e) [2 pts] In the primal version of SVM, we are minimizing the Lagrangian with respect to w and in the dual version, we are minimizing the Lagrangian with respect to α. F

f) [2 pts] For the dual version of soft margin SVM, the α_i's for support vectors satisfy $\alpha_i > C$. F

Question 4) [16 pts] A local startup is using machine learning to classifier customers into several categories. The startup trained a classifier on a set of 1,000 labeled customers. Since this classifier achieved 100% accuracy with the customers in this training set, the startup deployed the classifier into a product. To its surprise, the classifier achieved an accuracy of only 80% after the deployment. Since you are a machine learning expert, the startup hires you as a consultant to help resolve this situation.

a) [8 pts] What mistake did the startup make? Explain how this mistake could explain why the accuracy of the classifier went from 100% before the deployment to 80% after the deployment.

Did not split the dataset or used training set as the test set. The classifier overfitted the training set and resulted the lower accuracy with the new example.

b) [8 pts] Describe an approach to evaluate the classifier that should ensure that the accuracy observed after the deployment is close to the accuracy measured before the deployment.

Could have done cross-validation or regularization.

Question 5 [12 pts] Logistic regression.
a) [6 pts] Suppose that you trained a logistic regression model on some training data and the resulting weights are $w_0 = 1, w_1 = 2, w_3 = -3$. Assuming two classes (+ and -), a data point is predicted to belong to class + when $\sigma(w^T x + w_0) \geq 0.5$. Classify the following data points:
i) $(2,1)^T$
ii) $(1,2)^T$
iii) $(0.5,-1)^T$

$$(1 \quad 2 \quad -3)\begin{pmatrix}1\\2\\1\end{pmatrix} = 1(1) + 2(2) - 3(1) = 2 \Rightarrow +$$

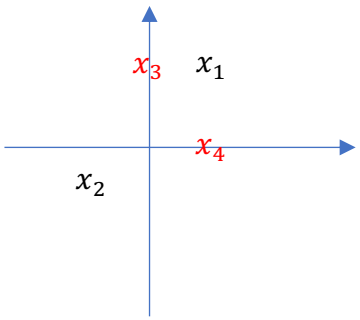
$$(1 \quad 2 \quad -3)\begin{pmatrix}1\\1\\2\end{pmatrix} = -3 \Rightarrow -$$

$$(1 \quad 2 \quad -3)\begin{pmatrix}1\\0.5\\-1\end{pmatrix} = 5 \Rightarrow +$$

b) [6 pts] Consider the following data set:

$$x_1 = (1,2)^T \quad y_1 = +$$
$$x_2 = (-1,-1)^T \quad y_2 = +$$
$$x_3 = (0,2)^T \quad y_3 = -$$
$$x_4 = (1,0)^T \quad y_4 = -$$

where the first two points belong to class + and the last two points belong to class −. Is it possible for a logistic regression classifier to correctly classify all points in this dataset? If yes, give weights that ensure correct classification? If no, explain why and describe an approach that could be used to modify the logistic regression classifier to correctly classify all those data points?



Cannot draw a linear separator to classify x_1 & x_2 and x_3 & x_4.

We can define a non-linear mapping Φ that maps the data into a new space that is linearly separable.