Name and Student ID:

Machine Learning BLG527E, Oct 31, 2013, 120mins, 9:30am-11:30am, Midterm Exam.
Name: **Number:**

1 30	2 20	3 20	4 15	5 15	Total 100

Signature:
Duration: 120 minutes.
Open books, closed notes. Write your answers neatly in the space provided for them. Write your name on each sheet. Good Luck!

QUESTIONS

QUESTION1) [30 points, 10 points each]

1a) *What are the differences and similarities between VC dimension and PAC bounds.*

VC Dimension h of a classification algorithm g is the maximum number of selected training instances which can be correctly labeled for all possible 2^h labellings. The VC dimension can be used to bound, with a certain probability, the test error in terms of the training error plus a term that depends on the number of training instances and the VC dimension.
PAC bound for a classification algorithm give us the number of training instances N needed if we require a certain test error with a certain probability.
Both VC dimension and PAC bounds are used to estimate the generalization/test error for a classifier.
In VC the selection of the location of instances is important. In PAC bounds training and test instances are required to be i.i.d. and from the same distribution.
Both bounds are very difficult to obtain for complex classification algorithms.

1b) *Assume you observed one coin tossed 11 times and the observations at time t=1 to t=11 are as follows: {H,T,H,H,H,T,T,T,H,T,T}. Is the outcome at toss t+1 independent of the outcome at toss t? Why or why not?*
Let X(t) and X(t-1) be the random variables which denote outcomes at time t and t-1 respectively. Based on the data, the joint probability table and the marginal probabilities are given in the table below.
X(t) and X(t-1) are independent iff P(X(t)|X(t-1))= P(X(t)) [You could also use P(X(t),X(t-1))= P(X(t))* P(X(t-1))]
P(X(t)=H)=4/10=2/5 P(X(t)=T)=6/10 =3/5
P(X(t)=H | X(t-1)=H)=2/5 = P(X(t)=H)=2/5
P(X(t)=T | X(t-1)=H)=3/5 = P(X(t)=T)=3/5
P(X(t)=H | X(t-1)=T)=2/5 = P(X(t)=H)=2/5
P(X(t)=T | X(t-1)=T)=3/5 = P(X(t)=T)=3/5
Therefore X(t) and X(t-1) are independent.

	X(t)=H	X(t)=T	P(X(t-1))
X(t-1)=H		0.3	0.5
X(t-1)=T	0.2	0.3	0.5
P(X(t))	0.4	0.6	

1c) *Assume that g is a linear model and for input x, g outputs the following:* $g(\mathrm{x},\mathrm{w}) = \mathrm{w}^{\mathrm{T}}\mathrm{x} + w_0$.

You need to make the parameters (w) of g take smaller values and hope that it will result in better generalization. Given a dataset $X = \{x^t, r^t\}_{t=1}^N$, *how would you obtain the solution for w and* w_0 *in this situation. (**Hint:** Modify the sum of squares error function and derive the solution for w and* w_0 *analytically.)*

For ease of notation, assume that we modify the input vectors and define a new parameter vector so that: x = [1 x] (i.e. we append a 1 to the beginning of each input vector x), $\mathrm{w} = [w_0\ \mathrm{w}]$. Then $g(\mathrm{x},\mathrm{w}) = \mathrm{w}^{\mathrm{T}}\mathrm{x}$. The modified sum of squares error function is: $E(w|X) = \frac{1}{2}\sum_{t=1}^{N}(g(\mathrm{x}^t,\mathrm{w}) - r^t)^2 + \frac{1}{2}\lambda\mathrm{w}^{\mathrm{T}}\mathrm{w}$

$$\frac{dE}{dw} = \sum_{t=1}^{N}(g(\mathrm{x}^t,\mathrm{w}) - r^t)\frac{dg(\mathrm{x},\mathrm{w})}{dw}\Big|_{x^t} + \lambda\mathrm{w} = \sum_{t=1}^{N}(\mathrm{w}^T\mathrm{x}^t - r^t)\ \mathrm{x}^t + \lambda\mathrm{w} = \sum_{t=1}^{N}\mathrm{x}^t\mathrm{x}^{t\mathrm{T}}\mathrm{w} + \lambda\mathrm{w} - \sum_{t=1}^{N}r^t\mathrm{x}^t = 0$$

Name and Student ID:

$$\left(\sum_{t=1}^{N} \mathrm{x}^t \mathrm{x}^{t\mathrm{T}} + \lambda \mathrm{I}\right) \mathrm{w} = \sum_{t=1}^{N} r^t \mathrm{x}^t \;\; then \;\; \mathrm{w} = \left(\sum_{t=1}^{N} \mathrm{x}^t \mathrm{x}^{t\mathrm{T}} + \lambda \mathrm{I}\right)^{-1} \left(\sum_{t=1}^{N} r^t \mathrm{x}^t\right)$$

QUESTION2) [20 points]

You are an expert fisherman and you know that the hamsi (a fish from Bosphorus) length is distributed according to a normal distribution with a certain unknown mean and standard deviation of 2cm. You have a prior belief that the mean is distributed according to another normal with mean of 10cm and standard deviation of 1cm. You go to the Boshphorus and catch N=10 hamsi whose lengths are (in cms) as follows:
12, 13, 14, 12, 11, 11, 12, 10, 11, 10

Based on the information above, what is your estimation of the hamsi length using:

3a) Maximum likelihood estimation

The maximum likelihood for the mean of the hamsi length distribution is=

$$m = \frac{1}{N}\sum_{t=1}^{N} \mathrm{x}^t = \frac{12+13+14+12+11+11+12+10+11+10}{10} = 11.6$$

3b) Bayes estimation

$x^t \sim \mathrm{N}(\theta, \sigma_o^2)$ and $\theta \sim \mathrm{N}(\mu, \sigma^2)$

for prior distribution: $\mu = 10, \sigma = 1$

for length distribution: $\sigma 0 = 2$

$$E[\theta \mid \mathcal{X}] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$$

$$E[\Theta|X] = \frac{10/4}{10/4+1/1} 11.6 + \frac{1/1}{10/4+1/1} 10 = 0.714 * 11.6 + 0.285 * 10 = 11.14$$

Name and Student ID:

QUESTION 3) [20 points]
Schizophrenia is a mental disorder where the patients confuse what is real and what is their imagination. It highly reduces the quality of life for both the patients and for people close to them. It is known that 1 in every 100 individual has schizophrenia. Recently scientists came up with a genetic signature that only 7 percent of schizophrenic people don't have it. And only 9 percent of healthy individuals have the signature. There is a gene therapy when the fetus is not more than 5 months old but it has risks of course. Even when the fetus will really have schizophrenia the cost of the procedure is decided to be 10 units. But if the fetus will not be schizophrenic the cost of the therapy is 60 units. If we do not apply the therapy and the fetus will have schizophrenia the cost is assumed to be 100 units. Suppose you are faced with a fetus that shows the genetic signature for schizophrenia, would you apply the therapy? Justify your decision.

We should choose the action with minimum risk. Let α_1 denote applying the therapy and α_2 denote not doing anything. Let S and T denote random variables taking values from the set {0,1}. S=1 when individual has schizophrenia and S=0 otherwise. Similarly, T=1 when the genetic signature exists and T=0 when it is absent. What is given is:

T=1
P(S=1) = 0.01
P(T=0 | S=1) = 0.07 => P(T=1 | S=1) = 0.93
P(T=1 | S=0) = 0.09 => P(T=0 | S=0) = 0.91

Loss matrix $\Lambda = \begin{bmatrix} 10 & 60 \\ 100 & 0 \end{bmatrix}$

We can calculate the risks of actions as follows:

$$R(\alpha_1|T=1) = \lambda_{11}P(S=1|T=1) + \lambda_{12}P(S=0|T=1)$$
$$R(\alpha_2|T=1) = \lambda_{21}P(S=1|T=1) + \lambda_{22}P(S=0|T=1)$$

We need to find posterior probabilities P(S=1|T=1) and P(S=0|T=1). Using Bayes formula:

$$P(S=1|T=1) = \frac{P(T=1|S=1)\,P(S=1)}{P(T=1)}$$

Marginal probability P(T=1) can be calculated as integrating over conditional probabilities for all possibilities of S.

$$P(T=1) = P(T=1|S=1)P(S=1) + P(T=1|S=0)P(S=0)$$
$$P(T=1) = 0.93 \cdot 0.01 + 0.09 \cdot 0.99 = 0.0984$$

Then

$$P(S=1|T=1) = \frac{0.93 \cdot 0.01}{0.0984} \cong 0.095 \quad => \quad P(S=0|T=1) = 1 - 0.095 = 0.905$$

Now we have all the information to calculate the risks associated with actions α_1 and α_2.

$$R(\alpha_1|T=1) = 10 \cdot 0.095 + 60 \cdot 0.905 = 55.25$$
$$R(\alpha_2|T=1) = 100 \cdot 0.095 + 0 \cdot 0.905 = 9.5$$

Since risk of applying the therapy is higher, we shouldn't apply the therapy

Name and Student ID:

QUESTION 4) [15 points]

Suppose you became the head of the department in the future and students have complained that amount of cheese (kaşar) in the sandwiches of the canteen has significantly reduced. You know that according to the regulations the cheese should be 50grams. After the complaints you go to the canteen and buy 10 sandwiches and measure the amount of cheese in them as follows:

51	47	46	48	49	53	47	51	46	49

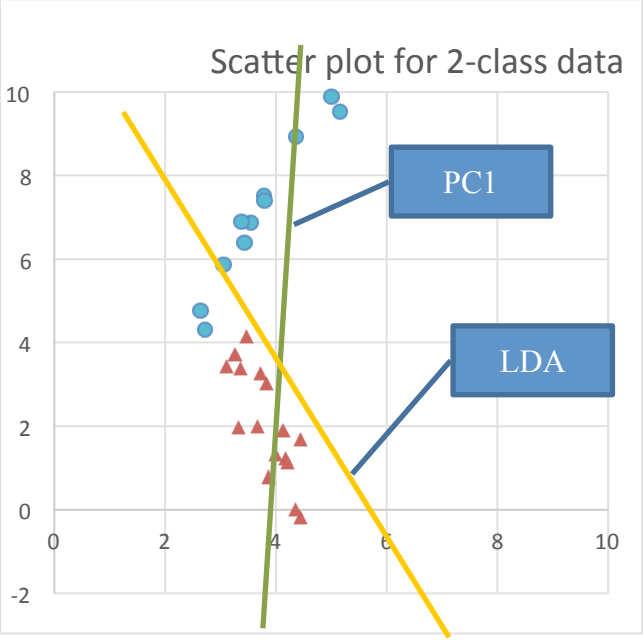
Can you decide with 95% confidence that amount of cheese is significantly reduced? Justify your decision. (**Hint:** $t_{0.05,9} = 1.83$, $t_{0.025,9} = 2.26$, $z_{0.05} = 1.64$, $z_{0.025} = 1.96$, choose wisely)

Since we don't know the variance and we only have 10 samples to calculate the sample variance, a *z*-distribution would be a very strong assumption and increases the risk of a Type I error (incorrect rejection of a true null hypothesis). So we better use a *t*-distribution which has a higher spread (heavy tails). We can calculate the *t*-statistic as:

$$t_{N-1} \sim \frac{\sqrt{N}(m-\mu)}{s} = \frac{\sqrt{10}(48.4-50)}{2.35} = -2.09$$

Since we are only interested in whether the amount is reduced or not (we don't care if the new mean is a lot higher than 50), we need to do a one-tailed *t*-test with 9 degrees of freedom ($t_{0.05,9}$). Since -2.09 < -1.83 we can reject the null hypothesis (that amount of kaşar did not change) with 95% confidence. We should warn the canteen.

QUESTION 5) [15 points]



Given the figure, assume you performed a PCA and LDA on this data. Draw the 1st principal component and LDA line on the figure. Explain the reasons why the two lines differ, if they do differ.

First principal component for PCA is always along the maximum variance for the whole data. PCA do not use the labels of the data.

On the other hand LDA line tries to maximize the distance between the projected means of the two classes while minimizing the sum within class variances. For this specific example, it can be said that projection on to PC1 would also result in bigger distance between the means but the LDA line is tilted because within class variances are high along PC1. On the LDA line, especially the variance of the blue class is significantly smaller.