Exam 03 February 2017
Duration: 2h30min

Name:_____

1. For a certain task, we collect a small sample of pairs (input t, output y), where t in the input and y is the output we desire to estimate:

t	y
5	722
10	1073
15	1178
20	1177

Assume a model $y = a + bt^2$.
a) Compute the values of a, and b with linear ridge regression, setting lambda = 0.5.
b) You are a bit worried since the model was obtained from a small sample. Luckily, yesterday you solved a similar/related task (but not equal) from a very large sample of 1000 observations and got a model $y_{related} = 600 + 2t^2$
How could you use this model to help you learn the new model? Can the learning task still be formulated as a standard (ridge) regression? If yes, then provide the solution.

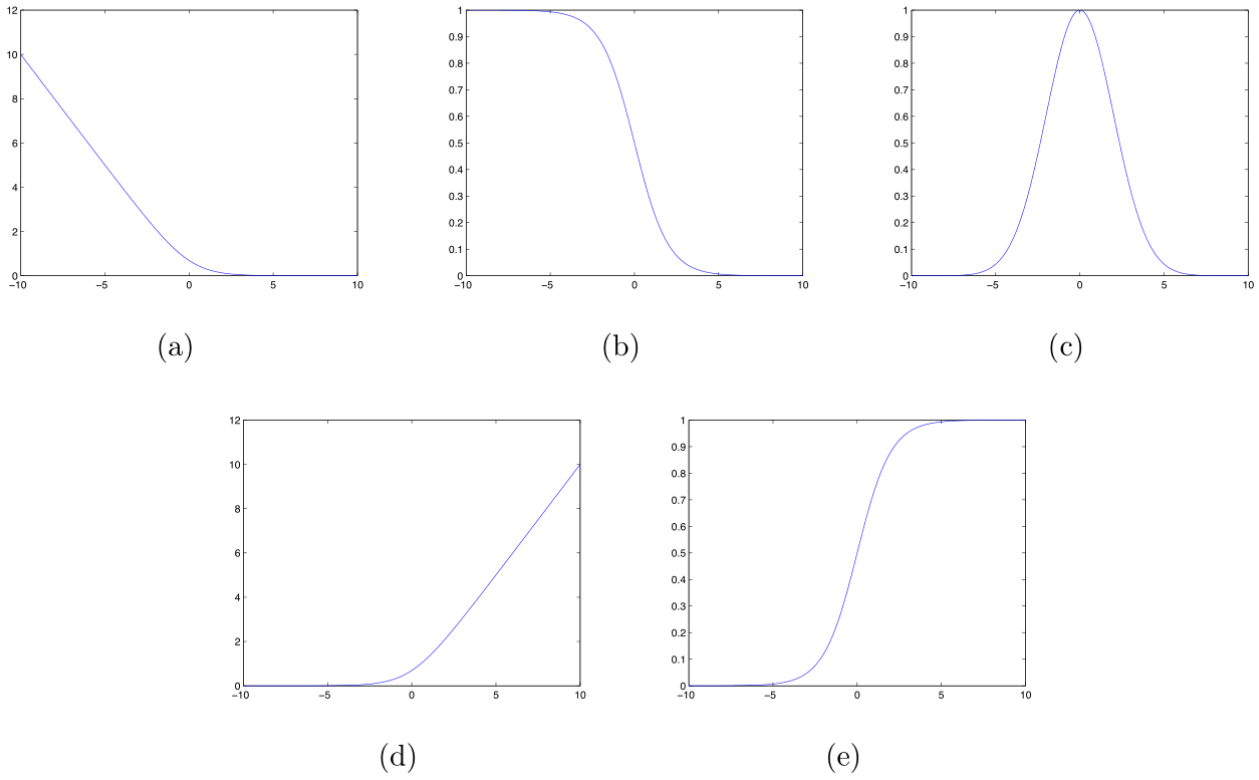
2. Several phenomena and concepts in real life applications are represented by angular data or, as is referred in the literature, directional data. Assume the directional variables are encoded as a periodic value in the range $[0, 2\pi]$.
Assume a two-class (y_0 and y_1), one dimensional classification task over a directional variable x, with equal a priori class probabilities.

a) If the class-conditional densities are defined as $p(x|y_0)= e^{2\cos(x-1)}/(2\,\pi\,2.2796)$ and $p(x|y_1)= e^{3\cos(x+0.9)}/(2\,\pi\,4.8808)$, what's the decision at x=0?
b) If the class-conditional densities are defined as $p(x|y_0)= e^{2\cos(x-1)}/(2\,\pi\,2.2796)$ and $p(x|y_1)= e^{3\cos(x-1)}/(2\,\pi\,4.8808)$, for what values of x is the prediction equal to y_0?
c) Assume the more generic class-conditional densities defined as $p(x|y_0)= e^{k_0\cos(x-\mu_0)}/(2\,\pi\,I(k_0))$ and $p(x|y_1)= e^{k_1\cos(x-\mu_1)}/(2\,\pi\,I(k_1))$. In these expressions, k_i and μ_i are constants and $I(k_i)$ is a constant that depends on k_i. Show that the posterior probability $p(y_0|x)$ can be written as $p(y_0|x) = 1/(1+e^{w_0+\,w_1\sin(x-\Theta)})$, where w_0, w_1 and Θ are parameters of the model (and depend on k_i , μ_i and $I(k_i)$).

3. Loss Function

Generally speaking, a classifier can be written as $H(x) = sign\,(F(x))$, where $H(x)$: $R^d \rightarrow \{-1,1\}$ and $F(x)$: $R^d \rightarrow R$. To obtain the parameters in $F(x)$, we need to minimize the loss function averaged over the training set: $\sum_i L(y^i F(x^i))$. Here L is a function of yF(x). For example, for linear classifiers, $F(x) = w_0 + \sum_{j=1}^{d} w_j x_j$ and $yF(x) = y(w_0 + \sum_{j=1}^{d} w_j x_j)$.

a) Which loss functions below are appropriate to use in classification? For the ones that are not appropriate, explain why not. In general, what conditions does L have to satisfy in order to be an appropriate loss function? **The x axis is yF(x), and the y axis is L(yF(x)).**



(a) (b) (c)

(d) (e)

b) Of the above loss functions appropriate to use in classification, which one is the most robust to outliers? Justify your answer.

c) Let $F(x) = w_0 + \sum_{j=1}^{d} w_j x_j$ and $L(yF(x)) = \frac{1}{1+\exp(yF(x))}$. Suppose you use gradient descent to obtain the optimal parameters w_0 and w_j. Give the update rules for these parameters.

Name:_____

4. Imagine a machine learning class where the probability that a student gets an "A" grade is P(A) =1/2, a "B" grade P(B) = μ, a "C" grade P(C) = 3μ, and a "D" grade P(D) = 1/2 − 4μ.

We are told that c students get a "C" and d students get a "D". We don't know how many students got exactly an "A" or exactly a "B". But we do know that h students got either an a or b. Therefore, a and b are unknown values where a + b = h. Our goal is to use expectation maximization to obtain a maximum likelihood estimate of μ.

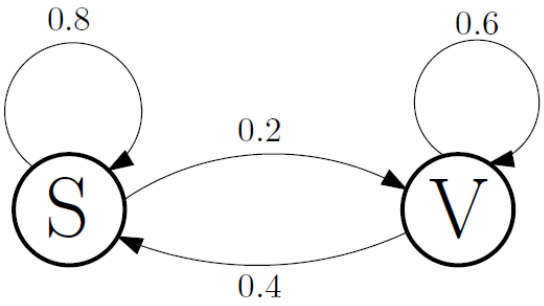
a) Expectation step: Which formulas compute the expected values of a and b given μ?

b) Maximization step: Given the expected values of a and b which formula computes the maximum likelihood estimate of μ? Hint: Compute the MLE of μ assuming unobserved variables are replaced by their expectation.

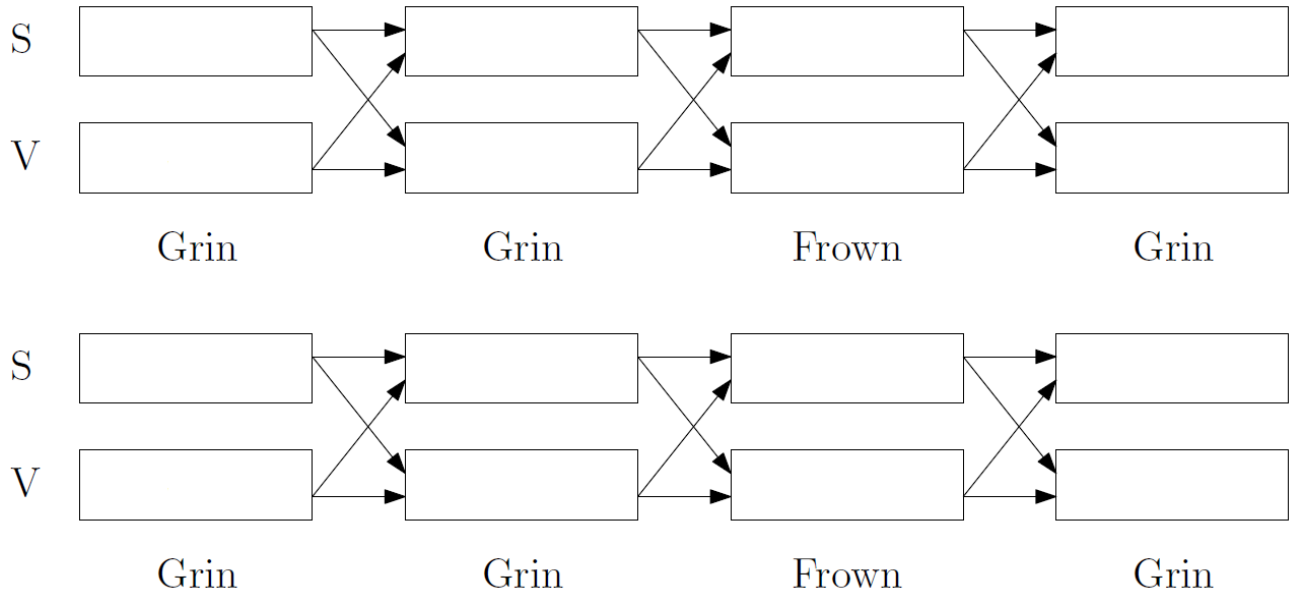
5. HMM. Consider the HMM below. In this world, every time step (say every few minutes), you can either be Studying or playing Video games. You're also either Grinning or Frowning while doing the activity. Suppose that we believe that the initial state distribution is 50=50.

E	$p(E\|X = S)$
grin	0.5
frown	0.5

E	$p(E\|X = V)$
grin	0.8
frown	0.2



a) We observe: **Grin, Grin, Frown, Grin**. What is the most likely path for this sequence of observations? Use the lattice below to provide intermediate results. (The lattice is repeated twice just for your convenience, in case you don't get it right at first)



b) Suppose that we **didn't know** the emission probabilities and transition probabilities for this HMM. Instead, we had to estimate them from data. Consider the following data set:

state: S S V V V S S S S S V S V V S V S S V V
Obs: G F G G F F F F G F G G G G F G F F G G

Based on this data, estimate the emission probabilities and the transition probabilities for this HMM.