Question 1 – Probability and estimation (12 points)

a) (4 points) Let a stick X_0 of unit length be broken at random at any position along the length with uniform probability. Let X_1 be the bigger piece. What is the expected length of X_1 ?

The stick is broken at position $x \sim U(0,1)$.

$\therefore$ Expected length $= E_x[\max(x, 1-x)]$

$$= \int_0^{1/2} (1-x) \cdot f_x(x)\,dx + \int_{1/2}^{1} x \cdot f_x(x)\,dx$$

$$= \int_0^{1/2} (1-x) \cdot 1 \cdot dx + \int_{1/2}^{1} x \cdot 1 \cdot dx$$

$$= \left[x - \frac{x^2}{2}\right]_0^{1/2} + \left[\frac{x^2}{2}\right]_{1/2}^{1}$$

$$= \frac{1}{2} - \frac{1}{8} + \frac{1}{2} - \frac{1}{8} = \frac{3}{4}$$

b) (4 points) Let a loss function be defined as the expected squared discrepancy between the actual and estimated value of the parameter, i.e., $g(\theta, \hat{\theta}) = E\left[(\theta - \hat{\theta})^2\right]$

and assume that we sample i.i.d. from the following distribution: $X_1, X_2, \ldots, X_n \sim N(\theta, 1)$.

Let $\hat{\theta_1} = X_1$ be an estimator which estimates the mean as the value of the first sample. What is the loss for this estimator?

$$g(\theta, \hat{\theta}_1) = E\left[(X_1 - \theta)^2\right] = \sigma^2$$ [where σ^2 = variance of the normal distrib., where $X_1 \sim N(\theta, \sigma^2)$ from defn.]

$$= 1$$ [$\because \sigma^2 = 1$]

c) (4 points) Let $\hat{\theta}_2 = \frac{\sum_i x_i}{n}$ be another estimator - the sample mean. What is the loss function for this estimator? [Hint: For any variable or function $E(y^2) = E(y)^2 + Var(y)$]

$$g(\theta, \hat{\theta}_2) = E\left[(M_n - \theta)^2\right]$$ (let M_n = sample mean on n samples

$$= [E(M_n - \theta)]^2 + Var(M_n - \theta)$$ (from corollary above

$$= 0 + Var(M_n - \theta)$$ ($\because$ sample mean is unbiased estimator of θ)

$$= Var(M_n) - Var(\theta)$$

$$= Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) - 0 = \frac{1}{n^2} \cdot n\, Var(X_1)$$ [$\because$ iid]

$$= \frac{1}{n} \cdot 1 = \frac{1}{n}$$

Question 2 – Naïve Bayes (16 points)

About 2/3 of your email is spam so you downloaded an open source spam filter based on word occurrences that uses the Naive Bayes classifier. Assume you collected the following regular and spam mails to train the classifier, and only three words are informative for this classification, i.e., each email is represented as a 3-dimensional binary vector whose components indicate whether the respective word is contained in the email.

‘study’	‘free’	‘money’	Category
1	0	0	Regular
0	0	1	Regular
1	0	0	Regular
1	1	0	Regular
0	1	0	Spam
0	1	0	Spam
0	1	0	Spam
0	1	0	Spam
0	1	1	Spam
0	1	1	Spam
0	1	1	Spam
0	1	1	Spam

a) (2 points) You find that the spam filter uses a prior p(spam) = 0.1. Explain (in one sentence) why this might be sensible.

Answer: It is worse for regular emails to be classified as spam than it is for spam email to be classified as regular email.

b) (4 points) Give the following model parameters when estimated as maximum-likelihood with add-one smoothing (i.e., using pseudocounts of one).

$P(\text{study}|\text{spam}) = 1/10$
$P(\text{study}|\text{regular}) = 2/3$
$P(\text{free}|\text{spam}) = 9/10$
$P(\text{free}|\text{regular}) = 1/3$
$P(\text{money}|\text{spam}) = 1/2$
$P(\text{money}|\text{regular}) = 1/3$

c) (5 points) Based on the prior and conditional probabilities above, give the model probability P(spam|***s***) that the sentence ***s***=“money for psychology study” is spam.

Answer: $p(\text{spam})\ P(\text{study}|\text{spam})\ (1-P(\text{free}|\text{spam}))\ (P(\text{money}|\text{spam})) =$
$p(\text{spam}) * 1/200$
$(1-p(\text{spam}))\ P(\text{study}|\text{regular})\ (1-P(\text{free}|\text{regular}))\ (P(\text{money}|\text{regular})) = (1-p(\text{spam}))*4/27$
$P(\text{spam}|s)= p(\text{spam})/200\ /\ [\ p(\text{spam})/200 +\ (1-p(\text{spam}))*4/27\ ]$
$= p(\text{spam})/200\ /\ [\ p(\text{spam})/200 + 4/27 - p(\text{spam})*4/27\ ]$
$= p(\text{spam})/200\ /\ [\ 4/27 - p(\text{spam})*773/5400\ ] = .003736$

d) (5 points) What should be the value of the prior p(spam) if we would like the above sentence to have the same probability as being spam as not spam, i.e., it would be classified as spam with probability 0.5?

$0.5 = p(\text{spam})/200\ /\ [\ 4/27 - p(\text{spam})*773/5400\ ]$ iff
$4/27 - p(\text{spam})*773/5400 = p(\text{spam})/100$ iff $p(\text{spam}) = .9674$

Question 3 – Regression (8 points)

We are dealing with samples x where x is a single value. We would like to test two alternative regression models:

1. y = ax + e
2. y = ax + bx^2 + e

We make the same assumptions we had in class about the distribution of e (e~N(0,s^2)).

a. (4 points) Assume we have n samples: x_1 … x_n . with their corresponding y values, : y_1 … y_n. Derive the value assigned to b in model 2. You can use a in the equation for b.

$$b = \frac{\sum_i y_i x_i^2 - a x_i^3}{\sum_i x_i^4}$$

b. (2 points) Which of the two models is more likely to fit the *training* data better?

   a. model 1
   b. model 2
   c. both will fit equally well
   d. impossible to tell

Answer: b. (model 2). Since it has more parameters it is likely to provide a better fit for the training data.

c. (2 points) Which of the two models is more likely to fit the *test* data better?

   a. model 1
   b. model 2
   c. both will fit equally well
   d. impossible to tell

Answer: d. It depends on the underlying model of the data and the amount of data available for training. If the data indeed comes from a linear model and we do not have a lot of data to train on model 2 will lead to overfitting and model 1 would do better. On the other hand if the data comes from an underlying quadratic model, model 2 would be better.

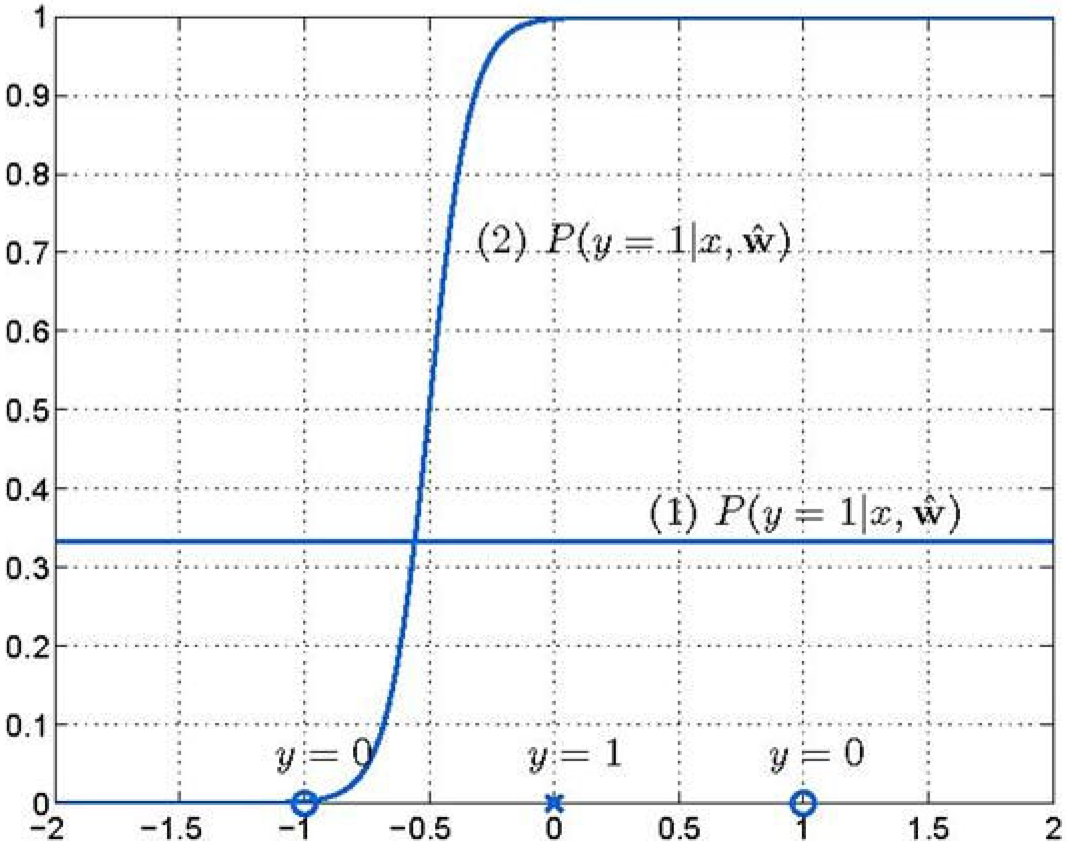
Question 4 – Logistic Regression (12 points)

Consider a simple one dimensional logistic regression model

$$P\ (y = 1|x,\ \mathrm{w}) = g(w_0 + w_1 x)$$

where g(z) = 1/(1+exp(z)) is the logistic function.

The following figure shows two possible conditional distributions P (y =1|x; w), viewed as a function of x, that we can get by changing the parameters w.



(a) (4 points) Please indicate the number of classification errors for each conditional given the labeled examples in the same figure.

Conditional (1) makes (1) classification errors

Conditional (2) makes (1) classification errors

(b) (4 points) One of the two classifiers corresponds to the maximum likelihood setting of the parameters w based on the labeled data in the figure, i.e. its parameters maximize the joint probability

P(y=0|x=-1; w) P(y=1|x=0; w) P(y=0|x=1; w)

Circle which one is the ML solution and briefly explain why you chose it:

Classifier 1 or Classifier 2

Answer: Class. 1 b/c it can't be classifier 2, for which P(y=0|x-1)=0

(c) (4 points) Would adding a regularization penalty $|w_1|^2/2$ to the log-likelihood estimation criterion affect your choice of solution (Y/N)? (Note that the penalty above only regularizes w_1, not w_0.)? Briefly explain why.

Answer: no, because w_1 is zero for Classifier 1, so no penalty is incurred. Therefore, if it was the ML solution before, it must still be the ML solution.

Question 5 – Decision Trees (14 points)

Decision trees

a. (2 points) What is the biggest advantage of decision trees when compared to logistic regression classifiers?

Answer: Decision trees do not assume independence of the input features and can thus encode complicated formulas related to relationship between these variables whereas logistic regression treats each feature independently.

b. (2 points) What is the biggest weakness of decision trees compared to logistic regression classifiers?

Answer: Decision trees are more likely to overfit the data since they can split on many different combination of features whereas in logistic regression we associate only one parameter with each feature.

For the next problem consider *n* two dimensional vectors ($x = \{x_1,x_2\}$) that can be classified using a regression classifier. That is, there exists a w such that

$$y = \begin{cases} +1 & \text{if} \quad w^Tx+b>0 \\ -1 & \text{if} \quad w^Tx+b£0 \end{cases}$$

c. (5 points) Can a decision correctly classify these vectors? If so, what is an upper bound on the depth of the corresponding decision tree (as tight as possible)? If not, why not?

Answer: Yes. One possible strategy is to split the points according to their x_1 values. Since the data is linearly separable, for each x_1 value there is a cutoff on x_2 so that values above it are in class 1 and below it in class -1. Splitting the data based on the x_1 values can be done with a log(n) depth tree and then we only need at most one more node to correctly classify all points so the total depth is *O(log n).*

d. (5 points) Now assume that these n inputs are not linearly separable (that is, no w exists for correctly classifying all inputs using linear regression classifier). Can a decision tree correctly classify these vectors? If so, what is an upper bound on the depth of the corresponding decision tree (as tight as possible)? If not, why not?

Answer: Yes. Similar to what we did for c we can split the points according to their x_1 values. However, since they are not linearly separable we cannot assume a cutoff on x_2 anymore. Instead, we may need to consider different values for x_2 again, the can be done in at most log(n) depth for a total of 2log(n) and an O(log n) depth.

Question 6 – Neural Networks (15 points)

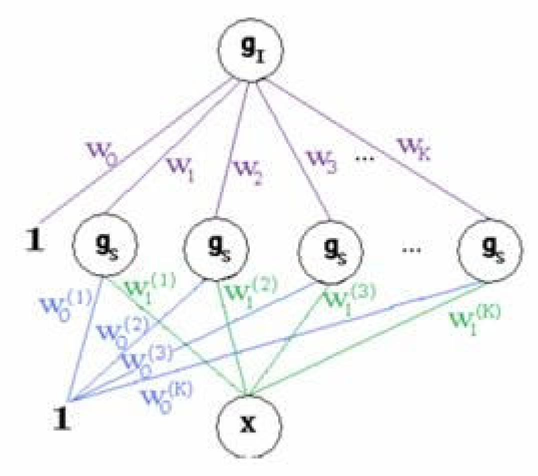
Suppose that you have two types of activation functions at hand:

- Identity function: $g_I(x)= x$
- Step function: $g_s(x)=1$ if $x >= 0$, 0 otherwise

So, for example, the output of a neural network with one input x, a single hidden layer with K units having step function activations, and a single output with identity activation can be written as

$$out(x) = g_I\,(w_0 + \Sigma_i\, w_i\, g_s\,(w_0^{(i)} + w_1^{(i)}\, x))$$

and can be drawn as follows:



1. (7 points) Consider the step function: u(x)=c if x<a, 0 otherwise (where a and c are fixed real-valued constants). Construct a neural network with one input x and one hidden layer whose response is u(x). Draw the structure of the neural network, specify the activation function for each unit (either g_I or g_S), and specify the values for all weights (in terms of a and c).
   Answer: g_I (c - $c*g_s(x-a)$)

   No points deduced for this (though not correct for x=a): g_I ($c*g_s(a-x)$)

2. (8 points) Now, construct a neural network with one input x and one hidden layer whose response for fixed real-valued constants a, b and c is
   c if x ϵ [a, b), and 0 otherwise.

   Draw the structure of the neural network, specify the activation function for each unit (either g_I or g_S), and specify the values for all weights (in terms of a, b, and c).

   Answer: g_I [$c*g_s$ (x-a) – c*g_s(x-b)]

Question 7 – Learning Theory (12 points)

Suppose you want to use a Boolean function to pick spam emails. Each email has n = 10 binary features (e.g. contains/ does not contain the keyword "sale").

a) Suppose the emails are generated by some unknown Boolean function of the n binary features (if the outcome of the boolean function is 1, it generates a spam otherwise regular emails).

Question:

i) (2 points) If our hypothesis space is all Boolean function, what is the error of the best hypothesis in our space?

Answer: 0.

Since the hypothesis space contains the true concept.

ii) (4 points) How many sample emails are sufficient to get a Boolean function with probability at least 95% that its error is less than 10%?

Answer: 7128

$$m \geq \frac{1}{\varepsilon}(\ln|H| + \ln\frac{1}{\delta})$$

$$\varepsilon = 0.1, \qquad \delta = 0.05, \qquad |H| = 2^{2^n} = 2^{1024}$$

$$m \geq 7127.8$$

b) Suppose the emails are generated by the following process: 75% of the emails are generated by Boolean function as described above. 25% of the emails are just random ones (randomly spam or not).

Question:

i) (2 points) If our hypothesis space is all Boolean function, what is the error of the best hypothesis in our space?

Answer: 12.5%

Since the accuracy will be 75% + 25% / 2 = 87.5%

ii) (4 points) How many sample emails are sufficient to get a Boolean function with probability at least 95% that its error is less than 15%?

Answer: 570223

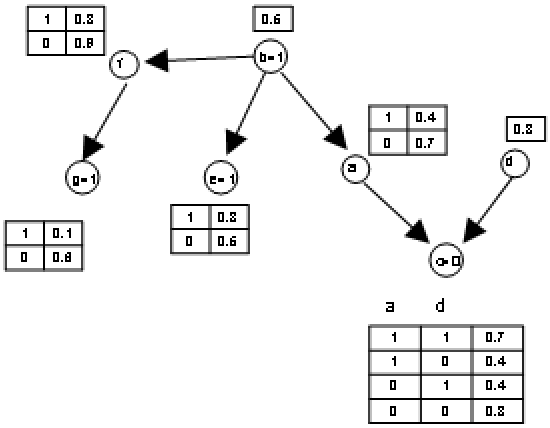
$$m \geq \frac{1}{2\varepsilon^2}(\ln|H| + \ln\frac{1}{\delta})$$

$$\varepsilon = 0.15 - 0.125 = 0.025, \qquad \delta = 0.05, \qquad |H| = 2^{2^n} = 2^{1024}$$

$$m \geq 570222.7$$

Question 8 – Bayesian Networks (11 points)

a. (6 points) The Bayesian network in the following figure contains both observed (denoted by 0 or 1 on the corresponding nodes) and unobserved variables.



All variables are binary. The CPTs for the probability of 1 for each variable are provided.
For example, p(c = 1| parent =1) = 0.3
p(c = 1| parent =0) = 0.9

Given the values observed, what is the value of p(a=1)?

Answer: Note that a is conditionally independent of all other nodes given its Markov blanket. Thus:

P(a=1|value of all nodes) = p(a=1|b,c,d)=

$$\frac{\sum_d p(a=1,b=1,c=0,d)}{\sum_d p(a=1,b=1,c=0,d)+\sum_d p(a=0,b=1,c=0,d)}$$

$$= \frac{0.5*0.4*(0.3*0.3+0.6*0.7)}{0.5*0.4*(0.3*0.3+0.6*0.7)+0.5*0.6*(0.6*0.3+0.2*0.7)} = 0.515$$

b. (5 points) Assume we are using Naïve Bayes for classifying an n dimensional vector

$x = \{x_1 \dots x_n\}$

that can belong to one of two classes (0 or 1). Draw the graph for corresponding Bayesian network. Note that there should be n+1 variables in your network. There is no need to provide any CPT.