6.867: Exam 1, Fall 2014

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page. Show your work neatly.

You may use any printed or written material or a calculator. You may not use any device with a wifi or cellular connection (even with the network turned off).

If a question seems vague or under-specified to you, make an assumption, write it down, and solve the problem given your assumption.

Come to the front to ask questions.

Name: _____ Athena ID: _____

Question	Points	Score
1	21	
2	6	
3	16	
4	12	
5	10	
6	12	
7	11	
8	12	
Total:	100	

Regression

1. (21 points) You are gathering data to estimate the brightness Y of an interesting astronomical object as a function of its position X. Assume both X and Y are real scalars. You have a cool new three-sensor array, so that for each position $x^{(i)}$ in your training data set, you have three brightness measurements, $y_1^{(i)}, \ldots y_3^{(i)}$. You assume that, given $x^{(i)}$, the $y_j^{(i)}$ values are independent and identically distributed; that is,

$$Y_j \sim \text{Normal}(w_0 + w_1 X, \beta^{-1}).$$

Suppose you have six brightness measurements from two observed positions, $x^{(1)}, x^{(2)}$. You and your friend Lars are both trying to estimate w_1 and w_0 using what you learned in 6.867.

(a) You decide to convert each observation $(x^{(i)}, y_1^{(i)}, y_2^{(i)}, y_3^{(i)})$ into three training examples: $((x^{(i)}, y_1^{(i)}), (x^{(i)}, y_2^{(i)}), (x^{(i)}, y_3^{(i)}))$ You use least-squares regression to estimate the w's by solving a system of the form

$$A^T B = A^T A W.$$

Write the entries of the design matrix A and the training matrix B for your regression, in terms of the original $x^{(i)}$'s and the $y_j^{(i)}$'s in the box below.

(b) What are your values for $A^T A$ and $A^T B$?

(c) Lars decides to first average the three-sensor array output, so for each i, he computes

$$d^{(i)} = \frac{1}{3} \sum_j y_j^{(i)}$$

He then makes that into a training example $(x^{(i)}, d^{(i)})$.
Lars has two averaged measurements from the same data that you were given, and uses least-squares regression to estimate the w's by solving a system of the form

$$C^T D = C^T C W.$$

Write the entries of the data matrix C and the training matrix D for Lars's regression, in terms of the original $x^{(i)}$'s the $y_j^{(i)}$'s in the box below.

(d) What are Lars's values for $C^T C$ and $C^T D$?

(e) Will your and Lars's W estimates be identical? If not, how would you adjust Lars's strategy to make them identical?

(f) You and Lars both decide that you have too little data and are worried about overfitting. You switch to finding W by solving a system of the form

$$A^T B = (A^T A + \lambda I)W$$

Lars, analogously, switches to solving

$$C^T D = (C^T C + \lambda I)W$$

Will your and Lars's W estimates be identical? If so, explain why. If not, how would you adjust Lars's strategy to make them identical?

(g) Finally, instead of doing maximum likelihood estimation, you and Lars decide to be Bayesian. You both start with a zero-mean Gaussian prior for the weights, with covariance $(1/\alpha)\mathrm{I}$. You treat A, B as your data and do a Bayesian update. What is your posterior covariance?

(h) Lars treats C, D as the data and does a Bayesian update and uses the same prior. What is his posterior covariance?

(i) Who is right? ◯ You ◯ Lars ◯ Both you and Lars ◯ Neither you nor Lars

Movies

2. (6 points) You have trained a predictor for the rating of a movie. Ratings are 1, 2, 3, 4, or 5 stars. Market research has shown that it is bad for your company to predict ratings that are too high. So, your loss function is this

	actual				
guess	1	2	3	4	5
1	0	1	1	1	1
2	2	0	1	1	1
3	3	2	0	1	1
4	4	3	2	0	1
5	5	4	3	2	0

What is the optimal prediction and its expected loss, for each of the following predicted distributions on the actual ratings?

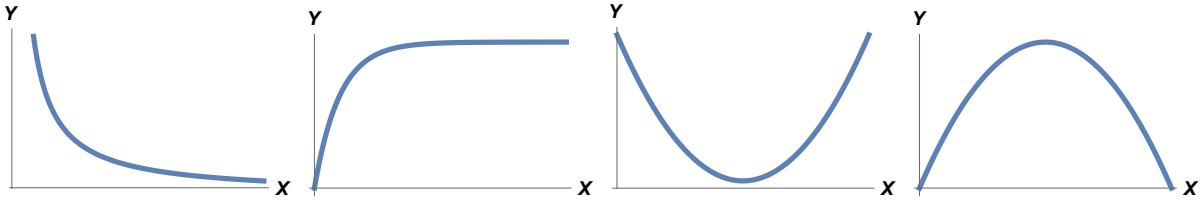
(a) $P(R)$ is uniformly distributed over $(1, 2, 3, 4, 5)$.

(b) $P(R = 5) = 0.9$, $P(R = 1) = 0.1$, all other probabilities 0

(c) $P(R = 5) = 0.5$, $P(R = 4) = 0.5$, all other probabilities 0

The Plot Thickens

3. (16 points) Consider just the general shape of the following plots. For each of the following possible interpretations of the quantities being plotted on the X and Y axes, indicate which of the plots could plausibly be the result. You may check any number of boxes (including 0) for each question.



(a) A (b) B (c) C (d) D

Assume all quantities other than X are held constant during the experiment. Error quantities reported are averages over the data set they are being reported on. Assume, in all cases but the very last, that the estimation procedure is some form of maximum likelihood estimation.

(a) **X**: Number of training examples **Y**: test set error
◯ A ◯ B ◯ C ◯ D

(b) **X**: Number of training examples **Y**: training error
◯ A ◯ B ◯ C ◯ D

(c) **X**: Number of training examples **Y**: cross validation error
◯ A ◯ B ◯ C ◯ D

(d) **X**: Order of polynomial feature set **Y**: test set error
◯ A ◯ B ◯ C ◯ D

(e) **X**: Order of polynomial feature set **Y**: training set error
◯ A ◯ B ◯ C ◯ D

(f) **X**: Order of polynomial feature set **Y**: cross validation error
◯ A ◯ B ◯ C ◯ D

(g) **X**: Number of training examples **Y**: variance in posterior
◯ A ◯ B ◯ C ◯ D

(h) **X**: Order of polynomial feature set **Y**: marginal likelihood
◯ A ◯ B ◯ C ◯ D

Model Selection

4. (12 points) Dana wants to use ridge regression to fit a model to data, use a model-selection procedure to select an appropriate value of λ, and produce a good estimate of how the resulting model will perform on unseen data. There are three data sets available: D_{train}, D_{validate}, D_{test}, all drawn from the same distribution. Define

$$J(w, \lambda, D) = \sum_i (y^{(i)} - w^T x^{(i)})^2 + \lambda \|w\|_2^2$$

Write the following answers in terms of J and the data sets.

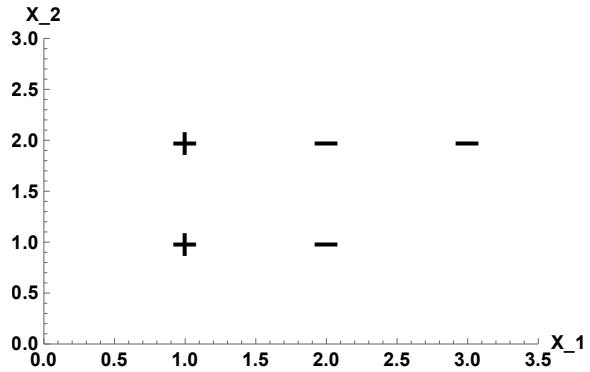
(a) Provide the definition of a function $W^*(\lambda)$ that specifies the optimal value of w for given a value of λ, according to the ridge regression criterion.

(b) Provide an expression for λ^*, which is the optimal value of λ. You may use the function W^* in your expression.

(c) Provide an expression for $\hat{E}$, the estimated error of the final resulting regression hypothesis on unseen data. You may use the function W^* and or the value λ^* in your definition.

SVMs

5. (10 points) Consider the following data set:



(a) What are the weights w_0, w_1, w_2 of the max margin solution

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

found by an SVM with slack variables as C tends to infinity? (If the solution is non-existent or not unique, explain).

(b) How many errors does it make on the training set?

(c) What are the weights w_1, w_2 of the max margin solution found by an SVM with slack variables as C tends to 0? (If the solution is non-existent or not unique, explain).

(d) What is the weight w_0 of the max margin solution found by an SVM with slack variables as C tends to 0? Recall that, in the soft-margin case, to compute w_0, we average over $\mathcal{M}$ which is the set of indices for which $0 < \alpha_i < C$:

$$w_0 = \frac{1}{|\mathcal{M}|} \sum_{j \in m} \left(y^{(j)} - \sum_{i \in S} \alpha_i y^{(i)} \left(x^{(j)^T} x^{(i)} \right) \right) .$$

If the solution is non-existent or not unique, explain.

(e) How many errors does it make on the training set?

Discriminant analysis

6. (12 points) You have scalar data drawn from two normal distributions. The class 1 data is drawn from a normal with $\mu_1 = 0$ and $\sigma_1 = 1$ and the class 2 data is drawn from a normal with $\mu_2 = 1$ and $\sigma_2 = 3$. Further, assume that the two classes are equally likely.

   (a) Suppose you had sufficient training data to accurately estimate the mean and variance parameters for the two normals, and then use the normals in an quadratic discriminant analysis procedure for classifying points. What range of data values would be placed in class 1? It is fine to write a condition involving x and constants.

   (b) Is there a linear classifier that will produce the same result? Why or why not?

(c) If your friend, who didn't take 6.867, accurately estimated the means for the two normals, but mis-estimated the variances as $\sigma_1 = \sigma_2 = 2$. In this case, What range of data values would be placed in class 1?

(d) Is there a linear classifier that will produce the same result? Why or why not?

Regular guys

7. (11 points) We are interested in regularizing the terms separately in logistic regression.

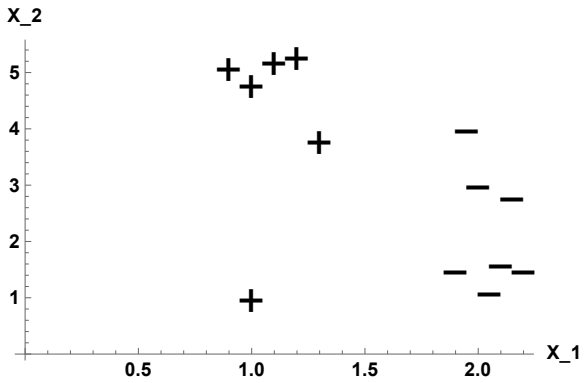
   (a) Consider the data in the figure below where we fit the model

   $$P(y = 1 \mid x, w) = \text{Sigmoid}(w_0 + w_1 x_1 + w_2 x_2)$$

   Suppose we fit the model by maximum likelihood, that is, we minimize

   $$J(w) = -\log \Pr(D_{\text{train}}; w)$$

   Sketch a possible decision boundary corresponding to w^*.



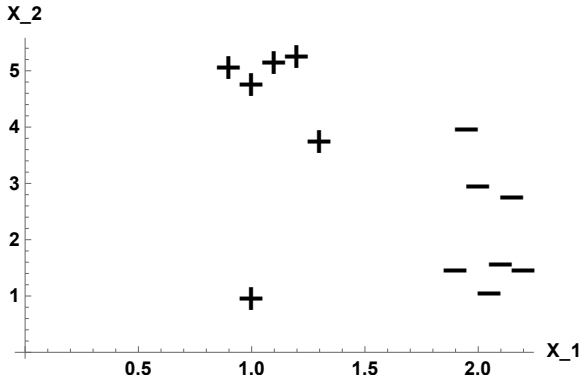
   (b) Is your decision boundary unique?

   (c) How many classification errors does it make on the training set?

(d) Now suppose we heavily regularize the w_0 parameter; that is, we minimize

$$J(w) = -\log \Pr(D_{\text{train}}; w) + \lambda w_0^2$$

with a very large value of λ. Sketch a possible decision boundary corresponding to w^*.

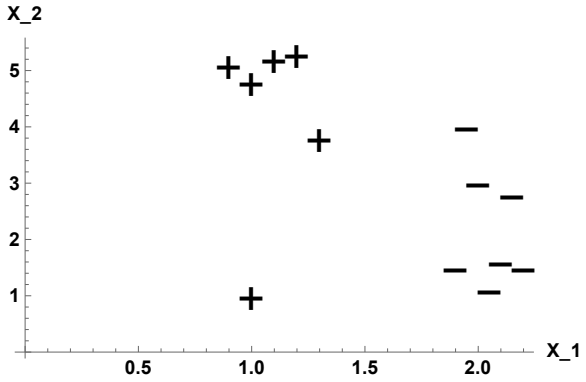


(e) How many classification errors does it make on the training set?

(f) Now suppose we heavily regularize only the w_1 parameter; that is, we minimize

$$J(w) = -\log \Pr(D_{\text{train}}; w) + \lambda w_1^2$$

Sketch a possible decision boundary corresponding to w^*.

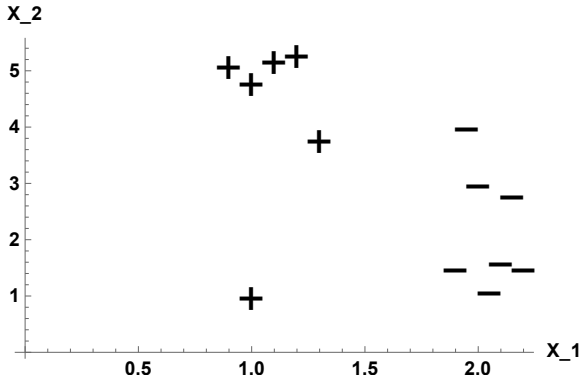


(g) How many classification errors does it make on the training set?

(h) Now suppose we heavily regularize only the w_2 parameter; that is, we minimize

$$J(w) = -\log \Pr(D_{\text{train}}; w) + \lambda w_2^2$$

Sketch a possible decision boundary corresponding to w^*.



(i) How many classification errors does it make on the training set?

Beta, loss and found

8. (12 points) Recall that, for binary variable Y distributed as a Bernoulli with probability P, and for data distributed according to a beta distribution $\text{Beta}(a, b)$. We will treat the event "Heads" as $Y = 1$ and "Tails" as $Y = 0$. Thus, the a parameter corresponds to "Heads" and the b parameter to "Tails."

   - The mean of P is $a/(a + b)$
   - The mode of P is $(a - 1)/(a + b - 2)$
   - The variance of P is $ab/(a + b)^2(a + b + 1)$
   - The probability that $\Pr(Y = 1)$ is $a/(a + b)$.

   Also, recall that *risk* is equal to expected loss.

   You find an ancient coin with a weird mass distribution, which leads you to have a $\text{Beta}(2, 4)$ prior on the outcome of a coin flip. You flip it four times and see data H, H, H, T and your loss function $L(g, a)$ is: $L(H, H) = 0, L(T, T) = 0, L(H, T) = 1, L(T, H) = 10$. You are asked to predict the next flip.

   (a) What is the risk of predicting H? (provide a number)

   (b) What is the risk of predicting T? (provide a number)

(c) Which of the following is true:

- ◯ You should prefer to predict heads.
- ◯ You should prefer to predict tails.
- ◯ You should be indifferent between heads and tails.

(d) Your friend Phred finds another ancient coin with a weird mass distribution. He flips it four times and gets the sequence H, H, H, T and then bets (with the same asymmetric loss function you used above) that the next flip will be T. Assuming Phred had a prior on the probability of heads Beta(a, b). What has to be true of the relationship between a and b in order for Phred's prediction to be optimal?