6.867: Exam 1, Fall 2015

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. You may not use any electronic device (calculator, phone, tablet, laptop, etc).

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

Write your name on every page.

Come to the front to ask questions.

Name: ______________________________ Athena ID: ______________________________

Question	Points	Score
1	24	
2	20	
3	18	
4	20	
5	24	
6	24	
Total:	130	

Name: ______________________________

Regularization

1. (24 points) You've collected data of the form $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ with $x^{(i)}, y^{(i)} \in \mathbb{R}$. You decide to use a conditional Gaussian model $Y \mid X = x \sim \mathcal{N}(w_0 + w_1 x, \sigma^2)$. For each of the expressions below, indicate whether it corresponds to: ordinary least-squares regression (OLS), MAP for Bayesian regression (BR), ridge regression (RR), LASSO, LAD, something else that is sensible (Good), or something else that is not sensible (Bad).

Only if it is one of these last two cases, explain why it is or is not a reasonable thing to do.

(a) $\operatorname{argmax}_w p(\mathcal{D} \mid w)$

○ OLS ○ BR ○ RR ○ LASSO ○ LAD ○ Good ○ Bad

(b) $\operatorname{argmax}_w p(w \mid \mathcal{D})$

○ OLS ○ BR ○ RR ○ LASSO ○ LAD ○ Good ○ Bad

(c) $\operatorname{argmax}_w p(\mathcal{D} \mid w)p(w)$

○ OLS ○ BR ○ RR ○ LASSO ○ LAD ○ Good ○ Bad

(d) $\operatorname{argmax}_w \prod_i p(x^{(i)} \mid y^{(i)}, w)$

○ OLS ○ BR ○ RR ○ LASSO ○ LAD ○ Good ○ Bad

(e) $\operatorname{argmax}_w \prod_i p(y^{(i)} \mid x^{(i)}, w)$

○ OLS ○ BR ○ RR ○ LASSO ○ LAD ○ Good ○ Bad

(f) $\operatorname{argmax}_w \prod_i p((x^{(i)}, y^{(i)}) \mid w)$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(g) $\operatorname{argmax}_w \sum_i p(y^{(i)} \mid x^{(i)}, w)$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(h) $\operatorname{argmax}_w \sum_i \log p(y^{(i)} \mid x^{(i)}, w)$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(i) $\operatorname{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(j) $\text{argmin}_w \sum_i |w_0 + w_1 \cdot x^{(i)} - y^{(i)}|$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(k) $\text{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2 - \lambda w \cdot w$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(l) $\text{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2 + \lambda w \cdot w$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(m) $\text{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2 + \lambda |w_0| + \lambda |w_1|$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

Name: ______________________________

Piecewise Linear Regression

2. (20 points) Suppose you were trying to do regression on a one-dimensional input space using a piecewise linear (but not necessarily continuous) function. A predictor with m pieces is parameterized with $m-1$ breakpoints $c_1, \ldots, c_{m-1}$ and m pairs $\beta_0^{(j)}, \beta^{(j)}$, so the regression function is

$$h(x) = \begin{cases} \beta^{(1)} \cdot x + \beta_0^{(1)} & \text{if } x \leqslant c_1 \\ \beta^{(2)} \cdot x + \beta_0^{(2)} & \text{if } c_1 < x \leqslant c_2 \\ \cdots & \\ \beta^{(m)} \cdot x + \beta_0^{(m)} & \text{if } c_{m-1} < x \end{cases} \tag{1}$$

The decision of how many pieces to use is part of the model-fitting process.

(a) If you were given 4 training points $\{(x^{(i)}, y^{(i)})\}_{i=1}^{4}$, give a set of parameters that would minimize sum squared error on the data. If it is useful, assume that $x^{(i)} < x^{(j)}$ for $i < j$.

(b) Is that set of parameters unique? Briefly explain why or why not.

(c) If you were required to limit yourself to predictors with $m = 2$, sketch an algorithm for finding the model parameters to minimize sum squared error on the data set with 4 training points.

(d) You are given 100 training examples, and you'd like to find a predictor (including m and parameter values) that you think will minimize expected squared loss on unseen data drawn from that same distribution. Sketch a procedure for doing this.

Name: ____________________

XOR

3. (18 points) Consider the following data set: $\{((0,0),-1),((1,0),+1),((0,1),+1),((1,1),-1)\}$.

   For each of the following feature representations, write down the corresponding kernel and specify whether or not it makes the data linearly separable. Draw the data in the new feature space and if it is linearly separable, draw in the separator that would be found by a hard-margin SVM.

   (a) $\phi(x) = (x_1 \cdot x_2, 1)$
   Kernel

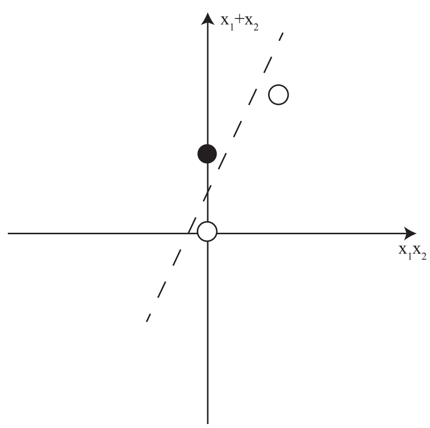
   Separable? ◯ True ◯ False

   The data points are not linearly separable. Because both $(0,0)$ and $(1,0)$ maps to the point $(0,1)$ in the feature space, but they have different labels.

(b) $\phi(x) = (x_1 \cdot x_2, x_1 + x_2)$
Kernel

Separable? ◯ True ◯ False

Data in feature space, separator if it exists (Black dots are positive, white dots are negative)

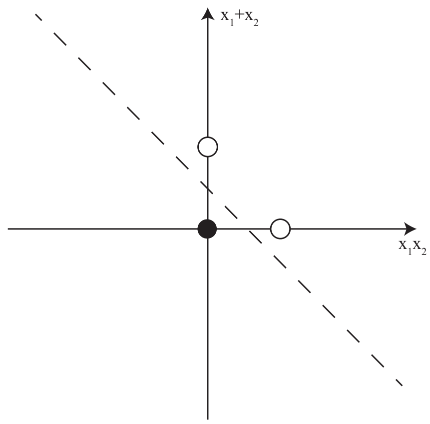


(c) $\phi(x) = (x_1 \cdot x_2, (1 - x_1) \cdot (1 - x_2))$
Kernel

Separable? ◯ True ◯ False

Data in feature space, separator if it exists (Black dots are positive, white dots are negative)

Name: ____________________



Name: ______________________________

SVM

4. (20 points) *Credit to Alex Smola*

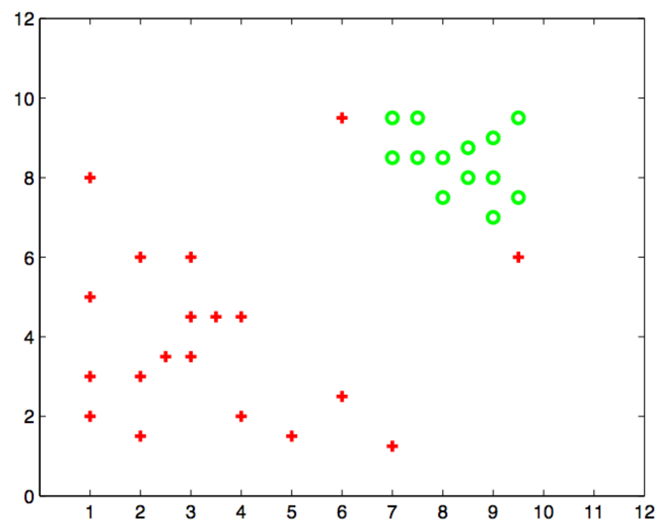
Recall that the primal form of the soft SVM optimization problem is

$$\min_{\theta,\theta_0} \left[\frac{1}{2}\|\theta\|^2 + C \sum_{i=1}^{n} \xi_i \right] \tag{2}$$

$$\text{s.t. } y^{(i)}(\theta^\top x^{(i)} + \theta_0) \geqslant 1 - \xi_i \tag{3}$$

$$\xi_i \geqslant 0 \tag{4}$$

Assume we are training an SVM with a **polynomial kernel** of degree 2. You are given the data set shown below. Please answer the following questions *qualitatively*, by giving a one sentence answer for each and drawing your solution, if one exists, on the given figure.



(a) Where would the decision boundary be if C is extremely large (i.e., as C goes to infinity)?

(b) For C near 0, where would the decision boundary be?

(c) In a classification task, which solution would work better: using a very large C or setting C to be near 0? Explain briefly.

(d) Draw a data point that *would not* change the decision boundary for very large C.

(e) Draw a data point that *would significantly* change the decision boundary for very large C.

Skee Ball

5. (24 points) Skee Ball is a carnival game, where a player tries to roll a ball up a ramp and get it to fall into a hole. Different holes win the player different numbers of points.

   Your skee ball game has three holes: a, b, and c.

   You can throw the ball soft (0) or hard (1).

   Initially, you don't know very much about how the game works; in particular, you don't know how your choice of throwing hard or soft affects which hole the ball falls into.

   So, you do some experiments!

   - You throw the ball soft 3 times, and it lands in a, a, and b.
   - You throw the ball hard 3 times and it lands in c, c, and c.

   Let H be the random variable indicating which hole the ball falls into (a, b, or c) and F be the random variable indicating how forcefully you throw the ball (0 or 1).

   For simplicity, we'll define $\theta_{hf} = P(H = h \mid F = f)$.

   In the all parts of this question, feel free to write out an expression with numbers plugged into it; you don't have to evaluate the expression numerically.

   (a) Having collected your experimental data, what are the maximum likelihood estimates of θ_{hf} for all values of $h \in \{a, b, c\}$ and $f \in \{0, 1\}$?

(b) We can think about the three parameters associated with a single conditional distribution as a point in a higher dimensional space: $\theta_f = (\theta_{af}, \theta_{bf}, \theta_{cf})$. Describe the set of valid values of θ_f?

(c) You want to be Bayesian and start with a uniform prior on θ_0 and a uniform prior on θ_1. What family of distributions, with what parameters, would you use for this purpose?

(d) What would the Bayesian posteriors on θ_0 and on θ_1 be, after conditioning on your experimental data? Provide distribution family (e.g. Gaussian) and numerical values (or detailed expressions) of the parameters.

Now assume that getting a ball into hole a is worth 1 point, into b is worth 5 points and into c is worth 4 points. We want to earn as many points as possible, and the loss relative to putting the ball into hole b (worth 5 points) is therefore $4, 0, 1$ for holes a, b, c.

(e) Let $\hat{\theta}_{hf}$ be the maximum likelihood estimate of getting a ball into hole h given how forcefully it was thrown. If we approximate θ_{hf} by using the MLE, write an expression for the approximate risk of each choice of how to throw the ball. What is the action that minimizes this approximate risk for the MLE calculated above?

(f) Assuming the Bayesian posterior is $p(\theta_f \mid \mathcal{D})$, write an expression for the posterior risk of each choice of how to throw the ball (i.e., write the risk where $p(\theta_f)$ is approximated by the posterior $p(\theta_f \mid \mathcal{D})$). What is the action that minimizes this approximate risk for the MLE calculated above?

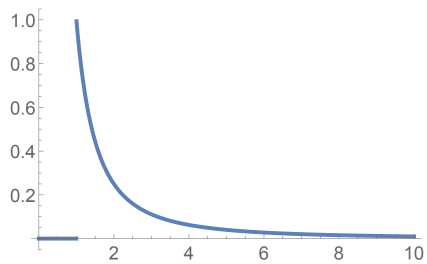
Name: ____________________

Pareto Optimal?

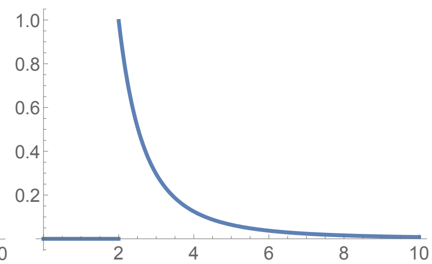
6. (24 points) You can get samples of a random variable X which is drawn uniformly at random from the interval $[0, M]$, but you don't know M. You model your prior belief on M using a Pareto distribution with parameters $1, 1$, which is shown in graph A below.

   A *Pareto* distribution has two parameters α and β both of which are real values greater than 0. Its pdf is

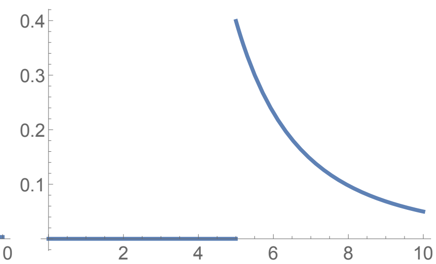
$$f_M(m) = \begin{cases} \frac{\alpha\beta^\alpha}{m^{\alpha+1}} & \text{if } m > \beta \\ 0 & \text{otherwise} \end{cases} \tag{5}$$



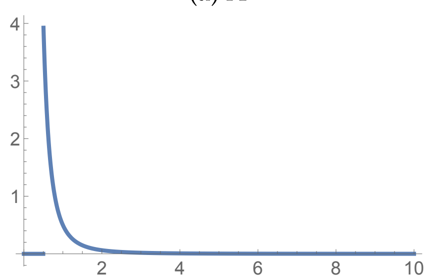
(a) A



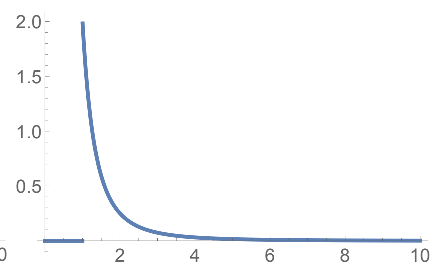
(b) B



(c) C



(d) D



(e) E

   (a) What is the pdf of the conditional distribution $p(M \mid X)$?
   (Hint: the Pareto distribution is a conjugate prior for uniform observations.)

   (b) If you start with a prior distribution $\text{Pareto}(1, 1)$ and observe $x^{(1)} = 0.5$, what is the family and parameters of the posterior distribution?

Name: ______________________

(c) Which of the graphs above does it correspond to?

◯ A ◯ B ◯ C ◯ D ◯ E

(d) If you start with a prior distribution Pareto$(1, 1)$ and observe $x^{(1)} = 5$, what is the family and parameters of the posterior distribution?

(e) Which of the graphs above does it correspond to?

◯ A ◯ B ◯ C ◯ D ◯ E