CS 189 Introduction to Machine Learning
Fall 2018 # Final Exam Common Errors

Before submitting a regrade request, please read the relevant sections of this document in their entirety. If you do submit a regrade request, specifically point to parts of your problem that were graded inconsistently with the rubric. Keep in mind that regrades will only be granted in cases of egregious or clear misgrading; arguments related to subjectively deserving more partial credit will most likely be dismissed.

Problem 1

(a) Grades are calculated as the maximum of (a) the number of rows that are *entirely* correct, and (b) the number of columns that are *entirely* correct.

(b) There were some creative answers here - nice! Many folks did not actually define random variables, or a proper distribution on their variables. Other than that many lost points because the variables they defined were indeed independent, or were actually correlated.

(c) We accepted "symmetric," and "can be written as $A^\top A$." Note that X does not need to be full rank, nor does it need to be positive semi definite.

(d) This one's pretty self explanatory.

(e) Points were allocated as follows: 2 points per complete part of the problem, where completeness requires providing the maximum value, as well as example u, v that achieve this value. Many people tried to argue about the covariance of the matrices; they aren't random vectors so this argument doesn't fit here.

Problem 2

(a) One of the common mistakes in this part was students not realizing kernel matrix was the identity. Some students forgot to denote what $\hat{f}(x)$. Some students only defined $\hat{f}(x)$ for $x = -1$ or $x = +1$ which is not correct as $\hat{f}(x)$ must be defined for all of $\mathbb{R}$

(b) One of the common mistakes in this part was students not realizing the kernel matrix was the matrix of all ones. Many students did not justify that $\alpha = 0$ was a minimizer to the function by showing it satisfies the normal equations for the given least squares problem. Many students tried to imply $\sum y_i = 0 \implies \|y\|_2 = 0$ which is not true.

(c) Common mistakes in this part was students not realizing the kernel matrix was $\begin{bmatrix} 1 + \frac{1}{2\sigma^2} & 1 - \frac{1}{2\sigma^2} \\ 1 - \frac{1}{2\sigma^2} & 1 + \frac{1}{2\sigma^2} \end{bmatrix}$. Many students miscalculated the inverse. Some students put 1s on the diagonal of this matrix. Many students also attempted to calculate $(KK^T)^{-1}$ which while is correct, involves much more calculation. Since K is PSD a calculation (K^{-1}) would have sufficed.

Problem 3

(a) A common answer consisted of saying that the objective is either non-convex or not strongly (strictly) convex, therefore it does not have a unique minizer. This argument is not valid. There exist functions that are non-convex and have a unique minimizer. Also, there exist functions which are convex and have a unique minizer, but they are not strongly convex. The confusion probably arose because: a strongly convex function has a unique minimizer. However, the converse is not true.

(b) Many people wrote the condition "$v_j >= 0$" for one of the cases of the derivative computation. However, v_j is defined to be maximum between zero and another quanitity. Therefore, it is always non-negative. If you wrote $v_j > 0$ as a condition and lost points, please submit a regrade request.

Many people used i to express the summation inside the indicator function. However, that i is different from the index of the parameter h_{ij}.

(c) As in the previous part, many people used i or j to index the summation over data points in the batch. This is incorrect since i and j index the parameter h_{ij}.

Many people did not explain how the batches of data points are selected for SGD.

Problem 4

(a) This part was looking for the accuracies in the population as a whole and in each subpopulation, where accuracy is defined as $P\{\widehat{Y} = Y\}$. Some students computed the positive predictive value due to interpreting 'recommendation' to mean $\widehat{Y} = 1$, so only one point was deducted in this case. Two points were deducted for computing the true positive rate instead.

(b) Some student mistakenly thought that accuracy, as computed in part (a), was related to the independence criterion. Recall that independence has to do with the event $\widehat{Y} = 1$, not the event $\widehat{Y} = Y$.

(c) Some students have answers that relied on being able to change the value of Y for an individual rather than only $\widehat{Y}$.

Problem 5

(a) -2: Giving heuristic explanation that lacks mathematical work/steps, lacks reference to linearity of expectation

(b) -3: Missing any relevant intermediate steps

(c) -3: Significant conceptual mistake. e.g. substituting $\mathbb{E}_S$ by a wrong sum, using large of large numbers/limits to argue equality, using generalization error wrongly (it's not 0 for a fixed vector)

(d) -1: Small mistakes

(b) -2: Giving heuristic explanation that lacks mathematical work/steps, lacks reference to linearity of expectation

(b) -3: Missing any relevant intermediate steps

(b) -3: Significant conceptual mistake. e.g. substituting $\mathbb{E}_S$ by a wrong sum, using large of large numbers/limits to argue equality, using generalization error wrongly (it's not 0 for a fixed vector)

(b) -1: Small mistakes

Problem 6

Common errors like "reverse proofs" and other things on the homework recieved varying deductions.

Part (a)

(a) Students lost 2 points for claiming that $\mathbf{V}_k\mathbf{V}_k^\top = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}$. This is false, trivializes the problem, and demonstrates a gross misunderstand about projection matrices.

(b) Students lost 1 point for numerous errors specifiying dimension, and not making sure dimensions worked out. Errors deemed more minor recieved only .5 point deductions. For example, if you used I_k and didn't define it, you may lost .5 points, because I_k typically is the $k \times k$ identity matrix, and typically this did not conform with problem dimensions.

(c) Students lost 1 point for not clearly specifying Σ

(d) If Σ_k was not given as a clearly diagonal matrix, 1.5 were deducted, because it meant that you did apply the requisite cancellations.

(e) Some sudents had $\langle \mathbf{v}_i, \mathbf{v}_i \rangle$ in their Σ. One point was deducted for this because to complete the answer, one needed to specify $\langle \mathbf{v}_i, \mathbf{v}_i \rangle = 1$.

(f) Some students used rank-one decompositions to prove the answer. This is totally fine, but you needed to convincingly show that appropriate cross terms canceled. Students whose solutions seemed to ignore the cross terms entirely lost 1 point.

(g) One point was deducted for not correctly answering that $\mathbf{X}\mathbf{P}_k = \mathbf{X}$. Some of you gave answers that correctly recognized that $\mathbf{\Sigma}_k = \mathbf{\Sigma}_d$ because $k \geq \text{rank}(X)$, but this did not receive points because you did not show that you recognized the conceptual point that $\mathbf{X}\mathbf{P}_k = \mathbf{X}$. If you said $\mathbf{X}\mathbf{P}_k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, then you only got .5 points off because this was closer to expressing $.\mathbf{X}\mathbf{P}_k = \mathbf{X}$.

(h) (Other Deductions) General minor proofs errors recieved one point deductions,
(Other Deductiions)

Part (b)

(b) There were a couple solution that were very difficult to parse, make numerous mistakes or little progress, but had some correct initial steps or ideas. These recieved 3 pt deductions.

(b) 1 point was deducted for not giving the correct ρ. Common errors here included not distinguishing between $i \leq k$ and $i > k$, not having that $\rho_{k,\lambda}(\sigma) = 0$ for $\sigma < \sigma_k$ (many students had that as $1/\lambda$), or using $\rho_{k,\lambda}(\sigma) = \mathbf{I}(\sigma \geq \sigma_k)\frac{\sigma}{\sigma^2+\lambda}$ (missing a factor of σ). If ρ was not specified,, 1.5 points were deducted.

(b) The problem said to consider *prediction*, not just the parameter. If you didn't touch the prediction part or complete, you lost 1.25 points. If you did do the prediction party, but wrote $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top$, you lost .75 points because $\mathbf{X}$ should have the normal Σ in is SVD.

(b) Students lost a variety of number of points based on number of missing cancellations, work not show, etc. This varied case by case

(b) Students tried to distributed inverses through sums, i..e $(A+B)^{-1} = A^{-1} + B^{-1}$. This is false, even for scalars, and even more wrong for matrices when one of the terms may not be (and isn't in this case bc of the PCA operation!) be invertible. 2 points were deducted

(b) Many students tried to solve the problem by analogy to Ridge regression. This is fine, but you needed to make clear what you were doing and not just "pattern match". Specifically, you needed to say that you are applying the ridge formula with $\mathbf{XP}_k$, and stating the corresponding singular values. Sometimes, I was lenient and allowed you to cite the ridge solution provided you only had a sum from $1:k$, but this typically required other indicators in the solution that you knew what you were doing. If this argument wasn't made with satisfactory clarity, you lost 1 point.

Part (c)

This was a difficult problem.

(c) Students who did not make any steps toward solving the problem lost 4 points, even if those steps were techinically correct.

(c) Students who started off writing out rank one decompositions lost 3 points because this was typically a good start.

(c) Students got far enough ad who tried to use the hint but didn't specify α_i lost 2 points, because its crucial to solving the problem. Students who also applied the inequality to $\alpha_i = \langle \mathbf{u}_i, \mathbf{y} \rangle$ lost 2 points, because this there was no bound on $\|\mathbf{y}\|$, only on W_*

(c) Students who wrote $\alpha_i = \mathbf{w}_i$ lost 1 point; the correct strategy was to use $\alpha_i = \langle \mathbf{v}_i, \mathbf{y} \rangle$.

(c) Many students tried to use Eckhart Young, which would be correct, but the students who tried this approach invariably confused the matrix operator norm and the vector L2 norm, which despite often similar notation, mean different things. These led to errors which garnered deuctiosn.

(c) Numerous students had various steps missing that led to varying deductions.

Problem 7

(a) The most common error was simply writing a derivative calculation with no mention of its purpose and/or no mention of convexity of the objective function. We were looking for the implication that convexity implied that setting the derivative equal to zero gave you a (unique, in this case) minimizer, and for you to carry out the calculation.

(b) There were many solutions given in the exam which demonstrated gaps in understanding of generalization error and risk. The grading rubric worked as follows. Recall that the problem was with 4 pts.

(i) 1 point was given for simply writing the average stability $\mathbb{E}[\epsilon_{\text{gen}}(\widetilde{w}_S, S)]$ in terms of the formulas given, e.g.

$$\mathbb{E}[\epsilon_{\text{gen}}(\widetilde{w}_S, S)] = \mathbb{E}_S \left[\mathbb{E}_{x \sim \mathcal{D}}[\text{loss}(\widetilde{w}_S, x)] - \frac{1}{n} \sum_{i=1}^{n} \text{loss}(\widetilde{w}_S, x_i) \right] ,$$

or

$$\mathbb{E}[\epsilon_{\text{gen}}(\widetilde{w}_S, S)] = \mathbb{E}_S \left[\mathbb{E}_{x\sim\mathcal{D}}[(1/2)(\widetilde{w}_S - x)^2] - \frac{1}{n}\sum_{i=1}^{n}(1/2)(\widetilde{w}_S - x_i)^2 \right] .$$

There were several ways students did this part incorrectly, and either received zero points or half a point.

- Some students treated $\widetilde{w}_S$ as a fixed vector that did not depend on S.
- Some students did not realize that the x appearing in $\mathbb{E}_{x\sim\mathcal{D}}$ was an independent copy of the n samples $x_1, ..., x_n$ represented by S, and instead treated x and x_i the same.
- Another mistake was to confuse the population distribution $\mathcal{D}$ from the empirical distribution over S. This manifested itself when students wrote $\mathbb{E}_{x\sim\mathcal{D}}[f(x)] = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$, for some $f(\cdot)$. Alternatively, other students assumed that $\mathcal{D}$ was a discrete distribution, but this assumption is not necessary to solve the problem.

(ii) Next, 2 points were given for taking the formula above, expanding the square correctly, and computing all the expectations correctly. Some students errornously claimed that $\mathbb{E}_S[R_S[\widetilde{w}_S]] = 0$, which is not true in general. Some students assumed that $\widetilde{w}_S = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean), which is not the case in this part of the question; $\widetilde{w}_S$ here is an arbitrary symmetric estimator. Some students did not realize that $\mathbb{E}[(x_i)^2] \neq (\mathbb{E}[x_i])^2$.

(iii) Finally, 1 point was given for correctly invoking the symmetry of the estimator $\widetilde{w}_S$. There were two ways to do this correctly. First, was to use symmetry to argue that $\mathbb{E}_S[\widetilde{w}_S x_i] = \mathbb{E}_S[\widetilde{w}_S x_1]$. The other way was to argue that $\mathbb{E}_S[\text{loss}(\widetilde{w}_S, x_i)] = \mathbb{E}_S[\text{loss}(\widetilde{w}_S, x_1)]$. Many students simply wrote $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[\widetilde{w}_S x_i] = \mathbb{E}_S[\widetilde{w}_S x_1]$ without any justification.

Several students confused generalization gap with stability. That is, some students tried to compute something similar to:

$$\mathbb{E}_{S,S'} \left[\frac{1}{n}\sum_{i=1}^{n} \text{loss}(\widetilde{w}_S, x_i) - \frac{1}{n}\sum_{i=1}^{n} \text{loss}(\widetilde{w}_{S'}, x'_i) \right] ,$$

or

$$\mathbb{E}_{S,S',x\sim\mathcal{D}} \left[\text{loss}(\widetilde{w}_S, x) - \text{loss}(\widetilde{w}_{S'}, x) \right] .$$

where S, S' differed by the introduction of one ghost sample. The quantity above is not the generalization error that we are looking at in this problem. Instead, the former equation measures the average sensitivity of the empirical risk to a one sample perturbation of the input, whereas the latter equation measures the average sensitivity of the generalization error to a one sample perturbation of the input.

(c) Common errors included:

(i) Correctly invoking part (b), but not stating that the sample mean was symmetric, which is a requirement to use part (b)

(ii) Correctly invoking part (b), but then messing up the expectation calculation in some way (usually, arriving at 0 or forgetting the $1/n$).