CS 189 Introduction to Machine Learning
Fall 2018

HW2 Common Errors

Common Problem 2 Errors

Part A

- Starting from the statement to be proved, and showing that this implies a true statement. In general, $A \implies$ TRUE does not imply that A itself is TRUE. In the context of problem 2(a), this means that starting from the statement $(AB+\mu I)^{-1}A = A(BA+\mu I)^{-1}$ and showing a series of forward steps that implies something true like $A = A$, is not enough. You would have to show that all of your steps are if and only if steps. In general, start with the true statement, and work your way to the statement you wish to probe. *We did not take off for this unless you explicity only showed one direction because so many people made this error, but in the future we will.*
- Only use the assumptions given. The only inverses you can assume are those given in the problem (A, B, AB, BA cannot be assumed invertible.)

Part B

- The easiest and least error prone way to answer this question was to derive the solution w^* from calculus as shown in class, and then to apply the result of part (a). Many submissions simply stated $(X^\top X + \lambda I)^{-1}X^\top = X^\top(XX^\top + \lambda I)^{-1}$ without stating that part (a) was being used; one point was deducted for this
- Many submissions left off at the calculus derivation, with $(X^\top X + \lambda I)^{-1}X^\top)$. Please read the problem statement carefully.
- Because a major part of this problem was invoking part (a), you need to specify why the conditions for part (a) hold, i.e. why we know $(X^\top X + \lambda I)^{-1}X^\top$ and $X^\top(XX^\top + \lambda I)^{-1}$ are invertible.

Part C

- One of the most common mistakes for this problem was in the dimensionality of Σ. The problem states that Σ is a diagonal matrix with entries $s_1, ..., s_d$, so that $\Sigma \in \mathbb{R}^{d\times d}$. Many submissions ignored this and chose to instantiate Σ from a different version of the SVD. When we explicitly provide you with the form, please use it.

Part D

- There were two major points we wanted to make here, (1) that the projection was onto the same space as in ordinary least squares, i.e. we are still projecting onto the columnspace of X, and (2) the singular values are smaller, so there is some shrinkage going on. Other insights than that were awesome. Simply saying “”we transform to a basis according to U, scale acoording to D, and transform back”, was not enough.

- Many people said P was a projection matrix. It is in general not, because the eigenvalues are not all 1.

Part E

- Many submissions provided no justification for math, but just a wall of equations. This means we can't give partial credit when there are mistakes. Not only is providing a description of what you're doing helpful for the reader, it also lets us know that you know what you're doing. *In general, a solution that consists of equations with no solutions might not get full credit, even if all equations are correct.*

- A specific instantiation of the above was when folks removed cross terms without stating why..

Part F

- When we ask for an interpretation of a quantity, please provide one. We can't give you credit if you don't write anything down.

- Many submissions gave an incorrect form for λ^*. Mostly these were algebra errors, so see solutions.

Part G

- Please note! This comparison only holds for the optimal λ^*, not arbitrary $\lambda > 0$.

- Many submissions did not compare to the OLS solution. This was explicitly a part of the problem, so in this case points were deducted.

- If you only compared to OLS in the limit, we took off 1 point for incomplete comparison. The comparison for any $\lambda^* > 0$ was necessary for full credt..

Common Problem 3 Errors

Part A

- The purpose of this problem was to realize that uncorrelation and independence are two separate things. While independence implies uncorrelation (part b), the converse is *not true*. Claiming that either non-independence implied correlation, or uncorrelation implied independence, resulted in a 3 points being lost.

- If a student correctly justified that the X and Y are independent, but skipped over correlation (or gave an incorrect answer that did not try to use correlation as a reason), only 2 points were deduced. Similarly, students who correctly computed that $\operatorname{Cov}(X, Y) = 0$ but skipped independence lost 2 points.

- We were pretty leninent with the correlation computation. Few people if any lost points for a calculation which showed no incorrect steps, and concluded that the covariance between X and Y was zero.

Part B

- The problem read "Verify that if X_i and X_j are pairwise independent, then Σ is diagonal, i.e. $\mathrm{Cov}(X_i, X_j) = 0$". Because we said *verify that*, at least some justification was required. Failure to provide any justification (e..g just saying independent implies zero covariance) resulted in a loss of two points.
- Gaussians are continuous random variables. If you tried to justify that independence implied zero covariance with sums, 1 point was deducted.

Part C

- First, some students used the "correlation"form of a bivariate Gaussian found on Wikipedia. Two points were deducted, unless you *explicitly established an equivalence between that form, and the form given to you in the problem statement*.
- A couple students argued that if X_1, X_2 are independent, then $\beta = 0$. This is correct, but the point of the problem was to verify *the opposite direction*. If you ared the "only if" using factorizatins, I was very lenient and only took off one bound. If you argued using that the covariance is zero, I took off many more points; see the next item.
- The point of Part a was to show that uncorrelation does not imply independence. If you argued that if $\beta = 0$, the covariance between X_1, X_2 is zero, and thus X_1, X_2 are independent, then you received no credit. Yes, it is true that for multivariate Guassians, these are equivalent, but *establishing this fact was precisely point of this problem!* If you just used uncorrelation to imply independence, you revealed to the grader that you do not understand the relationship between independence and correlation, and this grave conceptual error resulted in all points being deducted.
- Some students made computation errors computing the densities in terms of the covariance matrix provided. Between .5 and 2 points were deducted, based on the severity of the mistake.
- Some students claimed that when $\beta = 0$, the covariance matrix was the identity. 3 points were deducted for this.
- Lastly, some students did more work than they needed to. For examples, to show independence, you don't need to show that $f_{X_1}(x)$ nad $f_{X_2}(x)$ correspond to the marginals; that actually comes for free. Other students inverted Σ before setting $\beta = 0$, which was also more work than you needed to do.

Part D

- By far the most common deduction was taking the square root of the diagonal eigenvalues (usually labeled D or Λ), and not justifying why you could do this! Crucially, this relies on the fact that Σ has nonnegative eigenvalues, because its a covariance matrix (you didnt need to justify that, just say that its PSD).
- One more subtle point was that Σ is actually guaranteed to be PSD, not actually positive definite. To get that Σ is positive definite, you needed to use the invertibility of Σ, as stipulated in the problem statement. I only took of a tiny amount for this, but if you just said Σ is PD, or claimed that Σ^{-1} was PD without justification, I took of .25 points.

- Some students lost lots of points for very incorrect claims about which matrix properties imply what. Symmetric (or symmetric and invertible) does not imply PSD (e.g. -1 times the identity matrix); making this claim resulted in 2 points being deduced. Moreover, if you said symmetric and invertible directly implied a factorization, you lost 4 points.
- Students who just used a square root matrix without providing any justification lost points.
- Some students lost points by incorrectly reasoning about Σ instead of Σ^{-1}.
- Finally, some students used spectral decomposition without justifying where they came from, or giving matrices without explaining which terms corresponded to what. Lastly, if you refered to a spectral decomposition as an SVD, you lsot points.

Part E

- The question asked for the *minimum and maximum values*, not the minimizers and maximizer. Providing the minimizers/maximizers but not their values resulted in a loss of points.
- Many students tried to use theorems, which would be acceptable if you *correctly justified that the matrices in question satisfied the necessary conditions for those theorems to hold.* In particular, you need to note that you have a Raleigh quotient with the symmetric $A^\top A$ (not A!), and then cite that it is a standard fact. Note that A is not necessarily square, and for most students answer, it was not symmetric. Discussing the eigenvalues of A (which is not symmetric) lead to lots of points being lost.
- I took off .5 points for stating you answer in terms of the singular values of A rather than the eigenvalues of $A^\top A = \Sigma^{-1}$, or ideally, the eigenvalues of Σ. This was because in fact A is not unique, and the problem was given in terms of Σ.
- Students lost points for being unclear with what their eigenvalues, singular values and matrix decompositions were. Some students used incorrect decompositions and also lost points.
- A couple students only provided answers for $N = 2$, or assuming Σ was diagonal.
- Students overloading the notation Σ for the SVD of A and for the covariance matrix Σ in the problem. This is very confusing, and you lost a point for this.
- Some students made minor/moderate computational errors, including failing to compute reciprocals of eigenvalues, or taking unnecessary squares or square roots.

Part F

Students had a lot of trouble with this problem. Here are some errors

- By far the most troubling errors were students giving answers that did not parse.
- Many students did not read that x was constrained to the sphere and therefore incorrectly claimed that the minimizer was at $x = 0$.
- As in part (e), students provided minimizers and maximizers, but not minimum/maximum values.
- When Σ is diagonal, then the eigenvalues have interpreations in terms of variances. Referring to eigenvalues without mentioning variances resulted in loss of points.

- The correct answer for the first part of the problem was that the minimum value of $\|Ax\|_2^2$ was the reciprocal of the max variance of X_i (which is the last eigenvalue of Σ), and vice versa. You lost $.5$ points if you said the minimum value of $\|Ax\|_2^2$ *corresponded* to the max variance (this is not incorrect, but it is imprecise), 1.5 points if you said the minimum value of $\|Ax\|_2^2$ *was equal to* to the max variance (this is wrong), and 2 points if you just said that the minimum and maximum values corresponded to the miniumum and maximum variances, without distinguishing which was which.

- Lastly, students would write things like ”largest eigenvector” or use matrix decompositions which weren't explained. Not only is this poor justification, because dependening on which matrix you are talking about (Σ v.s. Σ^{-1}), the largest eigenvalue may be either correct or incorrect. Thus, it is impossible to assing proper credit.