CS 189 Introduction to Machine Learning
Fall 2018

Midterm 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.

Common Problem 1(a) Errors

This question was well done. People who got points taken off had some serious misconceptions about taking vector derivatives.

Some common errors:

- Wrote expressions that don't type check e.g. summing a scalar with a vector
- Missed a negative sign
- Treated x_i and w as scalars
- Thought that the full vector derivative was the sum of partial derivatives

1 point was also taken off for writing additional equations that were obviously wrong.

Common Problem 1(b) Errors

There were 2 key parts to this question:

- State or conclude that loss function is convex: 3 points
- Set gradient to zero to get $w^* = \bar{x}$: 2 points

If you said that the gradient is zero at $w^* = \bar{x}$, that only shows that w^* is a stationary point, but does not justify why it *minimizes* L. To show that w^* is the (global) minimum, you have to point out that L is convex, or equivalently, that its Hessian is Positive Definite everywhere.

Common mistakes:

- Arguments specifically only referred to local min/PSD in a local neighborhood. (Note that $\min L$ by definition refers to global min).
- Saying that the loss is non-negative and attains 0 at $w = \bar{x}$ (False!)
- Saying that gradient is lowest at $w^* = \bar{x}$ (False!)

1 point was taken off for making a wrong claim. Here is a list of common false statements. Some of these are appalling. If you made any of these errors, please make sure you convince yourself why they are false.

- the loss function L is linear
- the Hessian is always positive (should instead say positive definite. A positive matrix is a matrix in which all the elements are greater than zero, which is clearly not true.)
- the Hessian is 1 (should instead say I, or the identity matrix)
- the Hessian is the vector of all-ones
- the Hessian is 0
- the Hessian is + (that doesn't mean anything)
- the Hessian is convex (in this case the Hessian is a constant function so it is indeed convex, but saying that does not justify why the loss function is convex)
- the linear combination/linear transformation/affine transformation of convex functions is convex (False in general, should instead say positive linear combination or convex combination)
- you "assume" convexity (should instead say L *is* convex)
- Gradient is linear implies convex (False in general)
- Since it's a loss function, the Hessian is PSD (clearly untrue. Loss functions can be arbitrary, can even be non-convex)
- Tried to prove convexity of the squared norm but proof was clearly wrong/had a false step (e.g. wrong use of triangle inequality)
- $\|x_i - w\|^2$ is quadratic so it is convex (clearly untrue. Quadratic forms can be non-convex)
- Gradient is zero implies point must be a max or a min (it can also be a saddle point, which is neither a mean or a max)

Common Problem 2(a) Errors

There were three errors students most commonly made. First, students seemed to incorrectly recall or derive the ridge regression estimator as $(X^\top X - \lambda I)^{-1} X^\top y$, rather than its correct form, $(X^\top X + \lambda I)^{-1} X^\top y$. Note that subtracting off λI would lead to problems. For example, if $X^\top X$ has an eigenvalue equal to $\lambda > 0$, then $(X^\top X - \lambda I)$. May not be invertible.

Second, many students incorrectly computed the inverse of $(X^\top X + \lambda I)^{-1}$. Most students correctly simplified to $(V\Sigma^2 V^\top + \lambda I)^{-1}$, and even correctly pulled out the V to obtain $V(\Sigma^2 + \lambda I)^{-1} V^\top$. However, many students inocrrectly distributed the inverse, yielding $V(\Sigma^{-2} + \frac{1}{\lambda} I)V^\top$. To see why this is not correct, consider the scalar case. In this setting, we want to compute $(a^2 + b)^{-1}$. Distributing this inverse would imply $(a^2 + b)^{-1}$ is equal to $\frac{1}{a^2} + \frac{1}{b}$. This is false. For example, if say $a^2 = 1$, and b becomes very large, then $(a^2 + b)^{-1}$ tends to 0, whereas $\frac{1}{a^2} + \frac{1}{b}$ tends to $1 + 0 = 1$.

Lastly, many students expressed ρ as a matrix. ρ is supposed to be a scalar function of the singular values, as made clear in the problem statement.

Common Problem 2(b) Errors

Because the problem asked students to simply write down the prediction in the appropriate form, we could not in good faith deduct points for lack of justification. However, many students did provide justifications which were either incorrect, or confused the appropriate directions of logical implication, and these solutions recieved modest deductions. Specifically, students started by writing

$$\widehat{y}(x) = x^\top \sum_{i=1}^{d} v_i \rho(\sigma_i) u_i^\top y. \tag{1}$$

and then replaced $\rho(\sigma_i)$ with $\rho_\lambda(\sigma_i)$. This suggests that 1 is one step of your justification, which means that you are assuming Equation (1) in the problem statement to be true. The point of the problem was to verify that (1) is true, not just to assume (1) and pattern match. Only one point was deducted in this case. Here are some ways you would have received full credit:

1. Jumping straight to the answer, e.g. $\widehat{y}(x) = x^\top \sum_{i=1}^{d} v_i \rho_\lambda(\sigma_i) u_i^\top y.$ or $\widehat{y}(x) = x^\top \sum_{i=1}^{d} v_i \frac{\sigma}{\lambda+\sigma_i^2} u_i^\top y.$. If your answer for $\rho_\lambda(\cdot)$ in part (a) was incorrect, we *did not* double deduct.
2. Including some other correct justification, like writing out that $w_{\text{ridge}} = U \text{diag}(\rho_\lambda(\sigma_1), \ldots, \rho_\lambda(\sigma_d)) V^\top$.

Common Problem 2(c) Errors

In this problem, 2 points were awarded for arriving at the expression $\mathbf{w} = ((\mathbf{X}\mathbf{V}_k)^\top \mathbf{X}\mathbf{V}_k)^{-1} \mathbf{X}\mathbf{V}_k \mathbf{y}$ via reasoning about the OLS solution or deriving the optimum by setting the gradient equal to zero. An additional point was awarded for plugging in the SVD of $\mathbf{X}$. Many common errors arose at this stage:

- $\mathbf{X}\mathbf{V}_k$ is a rectangular matrix, so while it is full rank, it is not invertible.
- $\mathbf{V}_k \mathbf{V}_k^\top \neq \mathbf{I}$ because $\mathbf{V}_k$ is not square.
- $\mathbf{V}^\top \mathbf{V}_k = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{d-k \times k} \end{bmatrix}$, not identity. It is also not a square matrix with 0 along the diagonal, because then the dimensions would not be consistent.

One point was awarded for reasoning about $\mathbf{V}^\top \mathbf{V}_k$ based on the orthogonality of the columns without other major manipulation errors. A final point was awarded for arriving at the correct expression. Common incorrect expressions include:

- $\mathbf{w} = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \mathbf{U}^\top \mathbf{y}$, which is incorrect because then $\mathbf{w} \in \mathbb{R}^d$ instead of $\mathbb{R}^k$.
- $\mathbf{w} = \text{diag}(\sigma_1, \ldots, \sigma_k) \mathbf{U}^\top \mathbf{y}$, which is incorrect because the dimensions do not parse.
- $\mathbf{w} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i^\top \mathbf{y}$ which is incorrect because then $\mathbf{w} \in \mathbb{R}^1$.
- any $\mathbf{w}$ that depended on d singular values instead of only k singular values.

Common Problem 2(d) Errors

In this problem points were deducted for the following errors:

1. Most students forgot to featurize $\mathbf{x}$ into $\mathbf{V}_k^\top \mathbf{x}$, and wrote $\widehat{\mathbf{y}} = \mathbf{w}^\top \mathbf{x}$ instead of $\widehat{\mathbf{y}} = \mathbf{w}^\top \mathbf{V}_k^\top \mathbf{x}$. One point was deducted for this error.
2. One point was deducted for incorrect expressions for $\widehat{\mathbf{y}}$. Most commonly, this involved summing over d terms instead of k.
    - As a result of the mistake 1, the sum expression for $\widehat{\mathbf{y}}$ may have been missing $\mathbf{v}_i$ (resulting in inconsistent dimensions) or may have included an additional $\mathbf{e}_i$. No points were deducted in this case.
3. One point was deducted for and incorrectly defined expression $\rho_k(\sigma_i)$. Most commonly, the expression was missing the cases $\sigma_i \geq \sigma_k$ and $\sigma_i < \sigma_k$. Another common error was writing $\rho_k(\sigma_i) = \sigma_i$ rather than the reciprocal.
    - As a result of the mistake 1, some students included an additional term $\mathbf{v}_i^\top$ in their definition of $\rho(\sigma_i)$. No points were deducted in this case.
4. Two points were deducted if no expression for $\rho_k(\sigma_i)$ was specified.

Common Problem 2(e) Errors

The short answer is that in practice, we would choose hyperparameters with k-fold cross-validation (or holdout, or leave one out).

To get full credit, it needed to be clear how you portion the data into different sets, and how you use these sets to pick the best hyperparameters.

A common mistake was correctly describing cross validation or holdout but proposing to pick the best hyperparameter based on the *training error*. In these methods, you would pick the hyperparameter corresponding to the best average *validation error*.

Another common mistake was in responses that didn't explicitly mention cross validation or leave one out, assuming that there is a train/validate/test split. Youd need to instantiate those things from your data. This is part of the procedure were asking you to describe.

A lot of people chose to iteratively sweep over the hyperparameters to decide which hyperparameters to pick next. This is an ok solution to the problem, but its simpler to just define a grid search a priori.

A few people said to assign lambda to be the maximum singular value. First, note that this is an incomplete answer if it is not specified to be the maximum singular value of a certain matrix. Second, this might not be appropriate depending on the spread of the singular values.

Common Problem 3(a) Errors

Points were awarded only for every fully correct problem (one point per plot).

Separating lines had to go through the origin, see the problem statement where there is no bias term!

A few submissions just filled in the table, but without drawing the separating plane on the next page. These submissions received $1/2$ a point for each entry in the table, for a maximum of 2.5 points.

Common Problem 3(b) Errors

With $d = \sqrt{(x_1 - 5)^2 + (x_2 - 7)^2}$:

- Messing up the calculation if you used squared distance (e.g. $|d^2 - 9|$ or $|d^2 - 3|$)
- Not centering the classifier at $(5, 7)$
- Not using some sort of absolute value and only looking at distance from $(5, 7)$
- Using $|d - 2|$ or $|d - 4|$ as a classifier
- Decision boundaries or projection plot not corresponding to the given function
- Not plotting the data in the transformed space $\mathbb{R}$ (e.g. plotting it over radial distance from $(5, 7)$)
- Not giving a valid function definition (e.g. $\varphi(x) = 2 \leq d \leq 4$)
- Transforming the data to some space other than $\mathbb{R}$

Common Problem 3(c) Errors

- Saying that the max margin hyperplane is described by the line $y = -x$ or $y = -\frac{b}{a}x$. First of all do not use this notation because it can be confused with the notation for data points and labels. More importantly, the max margin hyperplane is given by the line $y = -\frac{a}{b}x$.

- Drawing the max margin separating hyperplane as going through the points $(a, -b)$ and $(-a, b)$. This is not the case. The max margin hyperplane goes through the points $(b, -a)$ and $(-b, a)$.

- Arguing that the hyperplane perpendicular to the vector $[a, b]^\top$ is optimal because the two points are equidistant to it. This is true, but not sufficient. Any hyperplane going through the origin is equidistant to the two points.

- Finding the optimal $w_\star$ and proving that it is optimal, but not saying that the max margin hyperplane is the plane going through the origin which is perpendicular to $w_\star$.

- Saying that the optimal w is equal to $[a, -b]^\top$ or $[b, a]^\top$ or other vectors that look similarly to the correct $[a, b]^\top$. Remember that the optimal w is perpendicular to the decision boundary.

Common Problem 4(a) Errors

- Saying that the second part of the problem follows directly from the first. It is a straightforward implication, but to get full credit you had to point out that you can define the kernels k_1 and k_2 so that they have the Gram matrices found in the first part.

- Giving a wrong example. For example, choosing $A_1 = A_2$ is not correct because $A_1 - A_2 = 0$ is a positive semidefinite matrix since all its eigenvalues are zero.

- Saying that a matrix is positive semidefinite because it is symmetric and all its entries are non-negative.

Common Problem 4(b) Errors

The intended solution to this problem was to explicitly construct a two point dataset $\{x_1, x_2\}$ where $x_1 \neq x_2$ (such as $\{x_1, x_2\} = \{0, 1\} \subseteq \mathbb{R}$) and then argue that the corresponding gram matrix was not positive semi-definite.

The rubric was as follows. You received 2pts if you wrote down a specific dataset where the corresponding gram matrix was not PSD. You then received an addition 3pts if you correctly argued why the gram matrix was not PSD, by either explicitly constructing a v such that $v^\top K v < 0$ or by explicit computation of the eigenvalues.

1 pt was given if you correctly identified that the diagonal entries of the gram matrix K would always be zero, but did not explicitly construct a dataset and just asserted this implied the gram matrix was not PSD.

We note that there was a different proof strategy than the one we intended which some students discovered. The proof was to assume towards a contradiction that k was valid. Then we know from Homework 3 by Cauchy-Schwarz that $k(x_i, x_j) \leq \sqrt{k(x_i, x_i)k(x_j, x_j)} = 0$. But then if we had any $x_i \neq x_j$, we would have $k(x_i, x_j) > 0$, which was the contradiction. This is a clever argument that received full credit. Some students tried to execute this argument but only looked at $k(x_i, x_i)$, which was not sufficient; incorrect execution of this argument received zero points.

The common errors on this problem are the following.

(a) One of the most common mistakes on this problem was to try to use 4(a) to "prove" that k was not a valid kernel. A typical "proof" of this flavor looked as follows. Students first expanded $k(x_1, x_2) = \|x_1\|_2^2 + \|x_2\|_2^2 - 2\langle x_1, x_2\rangle$. Students then argued that by defining $k_1(x_1, x_2) := \|x_1\|_2^2 + \|x_2\|_2^2$ and $k_2(x_1, x_2) := 2\langle x_1, x_2\rangle$, that both k_1 and k_2 were valid kernels and hence by 4(a) the difference $k_1 - k_2$ was not a valid kernel. This "proof" received zero points for the following reasons. First, k_1 is not a valid kernel. For instance, take the dataset $S = \{0, 1\}$ The gram matrix induced by k_1 is $\begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}$. The eigenvalues of this matrix are $1 \pm \sqrt{2}$ and hence this matrix is not PSD. More importantly, though, this logic is completely flawed. Just because $k_1 - k_2$ is not guaranteed to be a kernel, does *NOT* imply that $k_1 - k_2$ is always not a valid kernel. For instance, if we take $k_2(x_1, x_2) = 0$, this is a valid (trivial) kernel corresponding to the (trivial) feature map $\Phi(x) = 0$, and $k_1 - k_2 = k_1$ is a valid kernel! Some students also started from the expansion above and argued that $k_3(x_1, x_2) = \|x_1\|_2^2$ was a valid kernel and so was $k_4(x_1, x_2) = \|x_2\|_2^2$, and hence $k_3(x_1, x_2) + k_4(x_1, x_2)$ was a valid kernel. It is clear that both k_3, k_4 are *NOT* valid kernels– they both violate symmetry in the arguments (e.g. $k_3(x_1, x_2) \neq k_3(x_2, x_1)$).

(b) The next most common mistake was to write $k(x_1, x_2) = \langle x_1 - x_2, x_1 - x_2\rangle$ and then argue that it was obvious from this representation that one cannot write $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2)\rangle$ for some $\Phi(x)$. This is *NOT* a valid argument as it does not preclude the existence of a cleverly constructed $\Phi(x)$. This kind of argument also received zero points. Another common mistake was to write $k(x_1, x_2) = \langle x_1, x_1\rangle + \langle x_2, x_2\rangle - 2\langle x_1, x_2\rangle$ and then claim the $-2\langle x_1, x_2\rangle$ term was not necessarily ≥ 0 and hence the gram matrix was not PSD. This argument received zero points.

(c) Another common mistake was to claim that since the gram matrix had diagonal entries all equal to zero, it is not a PSD matrix. This is not true; consider the all zeros matrix. It has all diagonal entries equal to zero, but is a PSD matrix.

(d) Another mistake was that students claimed that in order for k to be a valid kernel we needed $k(x, x) = 0$

iff $x = 0$. This is not true; consider $x \in \mathbb{R}^2$ and the feature map $\Phi(x) = x_1$ (take the first coordinate of x). Here we have $k(x, x) = 0$ for $x = (0, 1)$.

(e) Some students tried to argue that a non-invertible gram matrix K implied that k was not a valid kernel. This is incorrect; kernel matrices do not necessarily have to be invertible, they just have to be PSD.

(f) Some students tried to aruge that it was enough to show that $k(x_i, x_j) < 0$. This is incorrect for two reasons. First, $k(x_i, x_j) \geq 0$ for our choice of k, since k is equal to the squared norm of the difference of x_i and x_j, and norms are always non-negative. Second, a valid kernel *CAN* have $k(x_i, x_j) < 0$ (consider the linear kernel $k(x_i, x_j) = \langle x_i, x_j \rangle$, for which the inner product most certaintly can be negative).

(g) Some students tried to argue that $k(x_i, x_j)$ was not symmetric. This is clearly false.