# CS 189 HW3: comments and common errors

October 9, 2018

**Legibility:** -2 for every part of question where there was a readability issue. Please latex, type or submit a high quality scan with good contrast + no glare or shadows + no distortion! use big, clear, dark, and legible handwriting. Make sure your hw scan is in the upright orientation, dont make your grader rotate it.

Your homework really should be readable beyond reasonable doubt. If the grader found it hard to read your solutions/had to squint, points were taken off this time.

**Problem 2a.**

- feature vector of $x_i$ is i-th eigenvector of $K = U\Lambda U^\top$ scaled by $\sqrt{\lambda_i}$, i.e. the column vector of $U\Lambda^{1/2}$. No. The feature vector of $x_i$ is the i-th row of $U\Lambda^{1/2}$. To see that the scaled eigenvectors are not the right choice note that the innerproduct between two distinct eigenvectors is zero while the off-diagonal entries of the Gram matrix might not be.
- answer contains variables that have not been defined. i.e. used $\Phi_X(x_i)$ without defining
- tried to use $v^\top Kv \geq 0$ characterization of PSD. This does not help you derive an explicit expression for the feature map in this question. remember that there are several (equivalent) characterizations of PSD matrices that you should try to use if one doesnt work.
- statement of answer is unclear, e.g. failed to define $\Phi_i$ as a row vector of $U\Lambda^{1/2}$. Its not obvious if $\Phi_i$ refers to column vector or row vector, and you wont always be given the benefit of the doubt for unclear presentation.
- many were confused and thought that the feature map is $\Phi x_i$ for some matrix $\Phi$. This is wrong. the f eature map is not (in general) a linear map!
- did not argue that $D^{1/2}$ or $\sqrt{\lambda_i}$ exists because eigenvalues are non-negative.
- did not justify existence of decomposition $K = U^T U$ (e.g. by referring to Cholesky or the spectral decompositions).
- $K$ has non-negative eigenvalues, not necessarily positive (that would be Positive Definite matrices).

- For matrices $\langle A, B\rangle := \mathbf{tr}(A^\top B)$ (matrix inner product) is NOT the same as AB (matrix product).

**Problem 2b.**

- if you only proved the inequality for a specific kernel, that doesnt answer the question.
- used wrong formula for determinant. Please make sure you know what the matrix determinant is.
- failed to say which kernel yields the C-S inequality. Need to say that youre applying $k(x_1, x_2) = x_1^\top x_2$.
- another mistake is saying $k(x_1, x_2) = |x_1^\top x_2|$. This is not a valid kernel. Taking absolute values of the entries of a PSD matrix does not yield a PSD matrix, in general.
- using Cauchy-Schwarz to show the first inequality. This answer also gets partial credit; note that when the question asked you to show C-S inequality, you shouldnt be using it. This includes arguing about the cosine of the angle between vectors (note that the cosine is only well-defined in general inner product spaces BECAUSE of the C-S inequality, so this is still circular reasoning.) Even if you attempt to prove C-S from first principles, this still does not get full credit if the proof only applies to finite-dimensional inner product spaces.
- plugging in $\Phi(x_i) = u$ and $\Phi(x_j) = v$. This does not define the feature map $\Phi$! You need to define the feature map at all $x \in \mathcal{X}$. This also does not convince the grader than you understand the Cauchy-Schwarz inequality; the inequality doesnt only apply to u and v (u and v are just arbitrary vectors in an inner product space).

**Problem 2c.**

- there were many algebra and indexing mistakes. Please read the solution carefully and study why/where you made a mistake in your computation.
- some people mistook matrix inner product (scalar) for the matrix product (matrix). For matrices $\langle A, B\rangle := Tr(A^\top B)$ (matrix inner product) is NOT the same as $A^\top B$ (matrix product) in general.
- showed that a feature map existed (e.g. by arguing K is PSD) without constructing one. When asked by a question to construct something, you must always state it as explicitly as you can.
- introduced new notation in an equation without defining it.
- saying $\Phi_3(x) = \Phi_1(x)\Phi_2(x)$. This is an invalid expression, $\Phi_1(x)$ and $\Phi_2(x)$ are (column) vectors of different lengths and one cannot multiply two column vectors.

- the definition of the feature map is vague. You need to define $\Phi_3(x) = \Phi_1(x)\Phi_2(x)^\top$ (i.e. define the matrix). If you defined every entry $\Phi_3(x)_{i,j} = \Phi_1(x)_i\Phi_2(x)_j$, defined $\Phi_3$ as a vector with the same entries or tried to vectorize $\Phi_3$, 0.5 points were taken off from your original number of points. Expressing $\Phi_3$ succinctly as an outer product of two vectors (and being aware of the matrix inner product) is a more elegant answer.

- indexing for vectors or matrices was poorly done, e.g. saying $\Phi_3(x) = [\Phi_1(x)_1\Phi_2(x)_1\Phi_1(x)_p\Phi_2(x)_q]$ and expecting the grader to infer the general pattern. You should be giving an unequivocal definition of $\Phi_3$.

- unlike in python, its not the convention to start indices from 0 in linear algebra. This is not a programming class—please dont use notation like $//$ and $\%$.

- points were also taken off for poor justification or poor presentation. Many answers were given without sufficient explanation or even without defining the notation (e.g. the index). If the grader had to mostly infer your reasoning, you wont be given full credit. It is also easy to make mistakes if you dont write out your steps clearly. Please refer to official solutions for a good example of how to answer the question.

- using terms like cross product, cartesian product to refer to the outer product is wrong. They are different concepts.

**Problem 3a.**

- did not normalize $\lambda$ (or the objective) by n. Either there needs to be a $1/n$ in front of your objective, or you need to have $\lambda \cdot n$ (instead of just $\lambda$) in your solution for $w$. People were given the benefit of the doubt.

- just copied the feature map from the notes without realizing the question wants you to show for p=2 (degree 2 polynomial)

- incomplete answer: didnt provide the features used for both types of regression.

- clarity of argument was key in this question. Solutions that proceeded in a step-wise fashion (computing the predictors for both kernel ridge regression and regularized polynomial regression, then justifying their equivalence) with ample explanation of what youre computing in each step received top scores.

- tried to show that the weights were equal instead of showing that the predictions were equal for the two types of regression.

- there is no need to use prove by contradiction for this problem. In general, using proof by contradiction for a problem where the solution is much more direct is not advised.

**Problem 3b.**

- computing the polynomia kernel takes time $O(\ell + \log p)$ not $O(\ell \log p)$ (here $\ell$ is the dimension of the given data points). To see this note that the polynomial kernel is $(1 + \langle \mathbf{x}_1, \mathbf{x}_2 \rangle)^p$. Therefore, after taking the innerproduct and adding one we get a scalar. There is no dependence on $\ell$ left.

- the matrix product between a $d \times n$ matirx $A$ and a $n \times d$ matrix $B$ takes time $O(d^2 n)$ operations, NOT $O(dn)$

- did not discuss the prediction complexity or said that polynomial regression and kernel regression have the same complexity for prediction or said that the prediction comlexity has a cubic dependence on the size of the problem. The parameters of these models need to be found only once and then reused for each new data point.

**Problem 3c.**

- many of you argued that the feature vector of each point can be a scalar because $e^{xy} = e^x e^y$. This is wrong; $e^x e^y = e^{x+y}$.

- you lost points for typos in the definition of the feature map because the whole problem was to get the feature map right.

- said that a linear combination of valid kernels is a valid kernels without specifying that the coefficients need to be non-negative. For example, in general $k_1 - k_2$ is not a valid kernel even if $k_1$ and $k_2$ are valid kernels.

- did not discuss the effect of $\gamma$ or said that $\gamma = 1$ recovers polynomial regression. When $\gamma = 1$ the factorials of the Taylor expansion still provide weighting to the features; which is not present in polynomial regression.

- wrote down the Taylor expansion as a sum over a finite number of terms.

**Problem 3d.**

- did not express correctly the minimum as an integral. For example, $\int_0^1 \mathbf{1}(t \geq x_i)\mathbf{1}(t \geq x_j)dt = 1 - \max\{x_i, x_j\} \neq \min\{x_i, x_j\}$.

- expressed the minimum as an integral of the form $\int_0^1 f_{i,j}(t) g_{i,j}(t) dt$, where $f$ and $g$ depend on both data points $x_i$ and $x_j$. This does not allow you to prove that the Gram matrices are PSD because the double summation with these functions cannot be written as a product of sums.

- proved that $2 \times 2$ Gram matrices are PSD. This is insufficient. Here is an example of a matrix whose principal minors are all PSD, but the matrix is not PSD

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

- said that a symmetric matrix with non-negative entries is PSD. This is wrong. For example, consider the matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which has eigenvalues $\pm 1$.

- proved that $\mathbf{x}^\top K \mathbf{x} = \sum_{i,j} x_i x_j k(x_i, x_j) \geq 0$. This is insufficient; you need to check that $v^\top K v \geq 0$ for any vector $v$ (not depending on the data points $x$). We did not deduct a lot of points for this because this was a difficult problem, but this is a serious mistake.

- said that $K$ is invertible because it is PSD. This is not true; a PSD matrix can have a zero eigenvalue and hence have the determinant equal to zero.

- said that you proved that $K$ is positive definite when you only proved that $K$ is PSD. To prove that $K$ is PD you would have to show that $v^\top K v > 0$ for all non-zero vectors $v$.

- said that $\int_0^1 Z(t)^2 dt = 0$ implies $Z(t) = 0$ for all $t \in (0, 1)$. This is not true. For example, if $Z(0.5) = 1$ and $Z(t) = 0$ for all $t \neq 0.5$, the integral of $Z(t)^2$ would still be equal to zero.

- said that $\alpha = K^{-1} y$ without offering an explicit formula for $\alpha_i$.

- NOTE that there is no need to show $K$ is invertible if you correctly find an explicit formula for $\alpha$ which solves the linear system. Since the labels $y_i$ are arbitrary, finding a solution of the linear system means that $K$ is invertible.

- gave an expression for $\alpha_i$ which depends on other coefficients $\alpha_j$. A solution to a system of equations which has a unique solution should be expressed only in terms of the coefficients of the equations.

- said that two scalars are *equivalent*. Statements can be *equivalent*; quantities can be *equal*.

**Problem 3e.**

- did not specify that the Gram matrix is invertible because it is positive definite or said that it is invertible because the data points are distinct. The data points need to be distinct because otherwise the Gram matrix cannot be positive definite.

- expressed kernel regression in terms of feature maps and used linear regression to argue on top of the feaures map to show that training error is zero. Such a proof is correct when the feature map is finite. In parts 3d and 3c we saw that feature maps can be infinite. How would you run linear regression with the feature map of the $\min$ kernel from part 3d? One can formalize this mathematically with more work, but it would still not be tractable to solve directly such a linear regression problem. Moreover, it is important to become comfortable thinking about the kernel regression as having coefficients $\alpha = K^{-1} y$ and making predictions

$$f(x) = \sum_i \alpha_i k(x_i, x).$$