CS 189 Introduction to Machine Learning
Spring 2018 HW6

1 Getting Started

Read through this page carefully. You may typeset your homework in latex or submit neatly handwritten/scanned solutions. Please start each question on a new page. Deliverables:

1. Submit a PDF of your writeup, **with an appendix for your code**, to assignment on Gradescope, “HW6 Write-Up”. If there are graphs, include those graphs in the correct sections. Do not simply reference your appendix.
2. If there is code, submit all code needed to reproduce your results, “HW6 Code”.
3. If there is a test set, submit your test set evaluation results, “HW6 Test Set”.

After you’ve submitted your homework, watch out for the self-grade form.

(a) Who else did you work with on this homework? In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

(b) Please copy the following statement and sign next to it. We just want to make it *extra* clear so that no one inadverdently cheats.

I certify that all solutions are entirely in my words and that I have not looked at another student’s solutions. I have credited all external sources in this write up.

This homework is due **Sunday, March 4 at 10pm.**

2 Canonical Correlation Analysis

The goal of canonical correlation analysis (CCA) is to find linear combinations that maximize the correlation of two random vectors. We are given two zero-mean random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$. Any linear combination of the coordinates of the random vector $\mathbf{X}$ can be written as $\mathbf{a}^\top \mathbf{X}$, and similarly, a linear combination of the coordinates of the random vector $\mathbf{Y}$ can be written as $\mathbf{b}^\top \mathbf{Y}$. Note that $\mathbf{a}^\top \mathbf{X}, \mathbf{b}^\top \mathbf{Y} \in \mathbb{R}$ are scalar random variables.

The goal of CCA can be summarized as solving the following optimization problem

$$\rho = \max_{\mathbf{a},\mathbf{b} \in \mathbb{R}^d} \rho(\mathbf{a}^\top \mathbf{X}, \mathbf{b}^\top \mathbf{Y}). \tag{1}$$

where the correlation coefficient $\rho(P, Q)$ between two zero-mean scalar random variables P and Q is defined by

$$\rho(P, Q) = \frac{\mathbb{E}[PQ]}{\sqrt{\mathbb{E}[P^2]\mathbb{E}[Q^2]}}.$$

The zero-mean jointly-Gaussian case expresses why we care about correlation. Two jointly Gausian zero-mean random variables that are uncorrelated are also independent of each other. Furthermore, if we want to best estimate (in a mean-squared sense) a particular scalar zero Gaussian random variable Y based on a vector $\mathbf{X}$ of jointly Gaussian zero-men random variables, then we are going to pick a weight vector $\mathbf{w}$ that is aligned with $\mathbb{E}[\mathbf{X}Y]$. It is straightforward algebra to see that this direction maximizes the correlation coefficient $\rho(\frac{\mathbf{w}^\top}{\|\mathbf{w}\|}\mathbf{X}, Y)$.

In this problem, we will work our way towards finding a solution to the problem of canonical correlation analysis using the singular value decomposition. In parts (a) to (c), we derive useful results regarding the singular value decomposition. In the later parts, you will use this result and see that canonical correlation analysis is equivalent to the singular value decomposition in a rotated and stretched coordinate system.

(a) Let $n \geq d$. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with full column-rank and singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we know that the singular values are given by the diagonal entries of $\mathbf{\Sigma}$, and the left singular vectors are the columns of the matrix $\mathbf{U}$ and the right singular vectors are the columns of the matrix $\mathbf{V}$. Note that both the matrices $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns.

Show that $\mathbf{A} = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where the ith singular value is denoted by $\sigma_i = \mathbf{\Sigma}_{ii}$ and $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith left and right singular vectors, respectively.

(b) **With the setup above, show that**

i) $\mathbf{A}^\top \mathbf{A}$ has i-th eigenvalue σ_i^2, with associated eigenvector $\mathbf{v}_i$.

ii) $\mathbf{A}\mathbf{A}^\top$ has i-th eigenvalue σ_i^2, with associated eigenvector $\mathbf{u}_i$.

Notice that both of the above matrices are symmetric.

(c) **Use the first part to show that**

$$\sigma_1(\mathbf{A}) = \max_{\substack{\mathbf{u}:\|\mathbf{u}\|_2=1 \\ \mathbf{v}:\|\mathbf{v}\|_2=1}} \mathbf{u}^\top \mathbf{A}\mathbf{v},$$

where $\sigma_1(\mathbf{A})$ is the maximum singular value of $\mathbf{A}$.

Additionally, show that if $\mathbf{A}$ has a unique maximum singular value, then the maximizers $(\mathbf{u}^*, \mathbf{v}^*)$ above are given by the first left and right singular vectors, respectively.

Hint 1: You may or may not find the following fact useful: We can express any $\mathbf{u} \in \mathbb{R}^n$: $\|\mathbf{u}\|_2 = 1$ as a linear combination of left singular vectors $\{\mathbf{u}_i\}$ of the matrix $\mathbf{A}$, and any vector $\mathbf{v} \in \mathbb{R}^d$ as a linear combination of the right singular vectors $\{\mathbf{v}_i\}$ of the matrix $\mathbf{A}$.

Hint 2: You may find the following results: For any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we have

- Cauchy-Schwarz inequality: $|\mathbf{a}^\top \mathbf{b}| \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$, with equality only when $\mathbf{b}$ is a scaled version of $\mathbf{a}$.
- Holder's inequality: $|\mathbf{a}^\top \mathbf{b}| \leq \|\mathbf{a}\|_1 \|\mathbf{b}\|_\infty$. Here, the ℓ_1 and ℓ_∞ norms of a vector $\mathbf{v}$ are defined by $\|\mathbf{v}\|_1 = \sum_i |v_i|$, and $\|\mathbf{v}\|_\infty = \max_i |v_i|$.
  - Let us say the vector $\mathbf{b}$ is fixed; then one way to achieve equality in the Holder inequality is to have: Let i be such that $|b_i| = \|\mathbf{b}\|_\infty$. Set $a_i = \|\mathbf{a}\|_1$, and $a_j = 0$ for all $j \neq i$.

(d) Let us now look at the canonical correlation analysis problem, where we are given the covariance of two zero-mean random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ as follows:

$$\mathbb{E}[\mathbf{X}\mathbf{X}^\top] = \mathbf{\Sigma}_{XX}, \quad \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = \mathbf{\Sigma}_{YY}, \quad \text{and}, \quad \mathbb{E}[\mathbf{X}\mathbf{Y}^\top] = \mathbf{\Sigma}_{XY}.$$

The goal is to find two directions/unit-vectors $\mathbf{a}$ and $\mathbf{b}$ so that the resulting scalar random variables $\mathbf{a}^\top \mathbf{X}$ and $\mathbf{b}^\top \mathbf{Y}$ have maximal correlation coefficient $\rho(\mathbf{a}^\top \mathbf{X}, \mathbf{b}^\top \mathbf{Y})$.

Show that the canonical correlation analysis problem can be rewritten as

$$\rho = \max_{\mathbf{a},\mathbf{b} \in \mathbb{R}^d} \frac{\mathbf{a}^\top \mathbf{\Sigma}_{XY} \mathbf{b}}{\left(\mathbf{a}^\top \mathbf{\Sigma}_{XX} \mathbf{a}\right)^{1/2} \left(\mathbf{b}^\top \mathbf{\Sigma}_{YY} \mathbf{b}\right)^{1/2}}.$$

Conclude that if $(\mathbf{a}^*, \mathbf{b}^*)$ is a maximizer above, then $(\alpha \mathbf{a}^*, \beta \mathbf{b}^*)$ is a maximizer for any $\alpha, \beta > 0$. We see that scaling the vectors $\mathbf{a}$ or $\mathbf{b}$ does not affect their correlation coefficient.

(e) Let us simplify our optimization problem. Assume that the covariance matrices $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ are full-rank. We choose matrices $\mathbf{\Sigma}_{XX}^{-\frac{1}{2}}$, $\mathbf{\Sigma}_{YY}^{-\frac{1}{2}}$ to whiten the vectors $\mathbf{X}$ and $\mathbf{Y}$ respectively. Note that for the vector $\tilde{\mathbf{X}} = \mathbf{\Sigma}_{XX}^{-\frac{1}{2}} \mathbf{X}$, we have

$$\mathbb{E}[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T] = \mathbb{E}[(\mathbf{X}\mathbf{\Sigma}_{XX}^{-\frac{1}{2}})^\top (\mathbf{X}\mathbf{\Sigma}_{XX}^{-\frac{1}{2}})] = \mathbf{I}$$

One may do a similar computation for the whitened vector $\tilde{\mathbf{Y}} = \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{Y}$, and conclude that its covariance matrix is identity too! Such a whitening step simplifies our computations, both at algebraic and conceptual level.

Rewrite the optimization problem from the previous part, using the aforementioned whitening in the following form:

$$\rho = \max_{\substack{\mathbf{c}:\|\mathbf{c}\|_2=1\\ \mathbf{d}:\|\mathbf{d}\|_2=1}} \mathbf{c}^\top \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{d}.$$

(f) Recall that the vectors $(\mathbf{a}^*, \mathbf{b}^*)$ denote maximizers in the previous part of the problem. Using the simplification and the above parts **show that**

ρ^2 is the maximum eigenvalue of the matrix

$$\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{XY}^\top\mathbf{\Sigma}_{XX}^{-1/2}.$$

Hint: An appropriate change of variables may make your life easier.

(g) Following the previous part's setup, **show that** $\mathbf{c}^* = \mathbf{\Sigma}_{XX}^{1/2}\mathbf{a}^*$ is an eigenvector corresponding to the maximum eigenvalue of the matrix

$$\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{XY}^\top\mathbf{\Sigma}_{XX}^{-1/2}.$$

(h) Following the previous part's setup, **show that** $\mathbf{d}^* = \mathbf{\Sigma}_{YY}^{1/2}\mathbf{b}^*$ is an eigenvector corresponding to the maximum eigenvalue of the matrix

$$\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{XY}^\top\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}.$$

(i) **Argue why such a CCA is not meaningful when the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, where by this we mean that $\mathsf{cov}(X_i, Y_j) = 0$ for all i, j.**

(j) Suppose you happen to know that $\mathbf{X}$ and $\mathbf{Y}^2$ (where a squared-vector $\mathbf{Y}^2$ is defined by squaring each entry of the vector $\mathbf{Y}$) share a linear relationship, **how could you still use CCA effectively with the given data?**

(k) **Why do you think that understanding CCA is relevant for machine learning?**

3 Mooney Reconstruction

In this problem, we will try to restore photos of celebrities from Mooney photos, which are binarized faces. In order to do this, we will leverage a large training set of grayscale faces and Mooney faces.

Producing a face reconstruction from a binarized counterpart is a challenging high dimensional problem, but we will show that we can learn to do so from data. In particular, using the power of Canonical Correlation Analysis (CCA), we will reduce the dimensionality of the problem by projecting the data into a subspace where the images are most correlated.

Images are famously redundant and well representable in lower-dimensional subspaces as the eigenfaces example in class showed. However, here our goal is to relate two different kinds of images to each other. Let's see what happens.

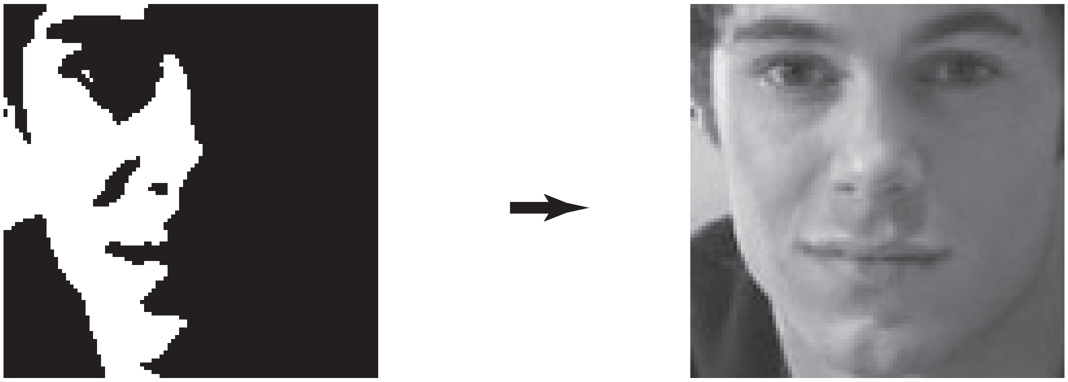


Figure 1: A binarized Mooney image of an face being restored to its original grayscale image.

The following datasets will be used for this project: X_train.p, Y_train.p, X_test.p and Y_test.p. The training data X_train.p contains 956 binarized images, where each image $\mathbf{X}_i \in \mathbb{R}^{15\times15\times3}$. The test data X_test.p contains 255 binarized images with the same dimensionality. Y_train.p contains 956 corresponding grayscale images, where $\mathbf{Y}_i \in \mathbb{R}^{15\times15\times3}$. Y_test.p contains 255 grayscale images that correspond to X_test.p.

Through out the problem we will assume that all data points are flattened and standardize as follows:

$$x \mapsto (x/255) * 2.0 - 1.0 \quad \text{and} \quad y \mapsto (y/255) * 2.0 - 1.0.$$

Note that this standardization step on loading the data does not ensure that the resulting images have zero mean. So removing the mean from the images is an additional processing step.)

Please use only the following libraries to solve the problem in Python 3.0:

```
import pickle
from scipy.linalg import eig
from scipy.linalg import sqrtm
from numpy.linalg import inv
from numpy.linalg import svd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
```

(a) We use CCA to find pairs of directions $\mathbf{a}$ and $\mathbf{b}$ that maximize the correlation between the projections $\tilde{\mathbf{x}} = \mathbf{a}^T\mathbf{x}$ and $\tilde{y} = \mathbf{b}^T\mathbf{y}$, From the previous problem, we know that this can be done by solving the problem

$$\rho = \max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^\top \mathbf{\Sigma}_{XY}\mathbf{b}}{\left(\mathbf{a}^\top \mathbf{\Sigma}_{XX}\mathbf{a}\right)^{1/2}\left(\mathbf{b}^\top \mathbf{\Sigma}_{YY}\mathbf{b}\right)^{1/2}},$$

where $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ denote the covariance matrices of the $\mathbf{X}$ and $\mathbf{Y}$ images, and $\mathbf{\Sigma}_{XY}$ denotes the covariance between $\mathbf{X}$ and $\mathbf{Y}$. Note that unlike in the previous problem, we are not given the covariance matrices and must estimate them from the images samples of $\mathbf{X}$ and $\mathbf{Y}$.

Write down how to estimate the three covariance matrices from finite samples of data and implement the code for it.

In order to properly compute the covariance matrices you have to 1) standardize the data, 2) subtract the mean and 3) scale by the number of data points at the end.

Note throughout the homework, we shall use the function StandardScaler to subtract the mean from the data and subsequently add it back.

(b) We know from the previous problem that we are interested in the maximum singular value of the matrix $\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}$, and that this corresponds to the maximum correlation coefficient ρ. Now, however, **plot the full "spectrum" of singular values of the matrix**

$$(\mathbf{\Sigma}_{XX} + \lambda\mathbf{I})^{-1/2}\mathbf{\Sigma}_{XY}(\mathbf{\Sigma}_{YY} + \lambda\mathbf{I})^{-1/2}.$$

For numerical issues we need to add a very small scalar times the identity matrix to the covariance terms. Set $\lambda = 0.00001$.

(c) You should have noticed from the previous part that we have some singular value decay. It therefore makes sense to only consider the top singular value(s), as in CCA. Let us now try to project our images $\mathbf{X}$ on to the subspace spanned by the top k singular vectors. Given the SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}$, we can use the left hand singular vectors to form a projection.

$$\mathbf{P}_k = \begin{bmatrix} \mathbf{u}_0 & \mathbf{u}_1 & \ldots & \mathbf{u}_{k-1} \end{bmatrix},$$

Show/visualize the "face" corresponding to the first singular vector $\mathbf{u}_0$. Use the following code for the visualization:

```
def plot_image(self,vector):
  vector = ((vector+1.0)/2.0)*255.0

  vector = np.reshape(vector,(15,15,3))

  p = vector.astype("uint8")

  p = cv2.resize(p,(100,100))
  count = 0

  cv2.imwrite('singular_face.png',p)
```

(d) We will now examine how well the projected data helps generalization when performing regression. You can think of CCA as a technique to help learn better features for the problem. We will use ridge regression regression to learn a mapping, $\mathbf{W} \in \mathbb{R}^{k\times 675}$ $(15 * 15 * 3 = 675)$, from the projected binarized data to the grayscale images. The binarized images are placed in matrix $\mathbf{X} \in \mathbb{R}^{956\times 675}$

$$\min_{\mathbf{W}} ||(\mathbf{X}\mathbf{P}_k)\mathbf{W} - \mathbf{Y}||_2^2 + \lambda||\mathbf{W}||_F^2.$$

Implement Ridge Regression with $\lambda = 0.00001$. Plot the Squared Euclidean test error for the following values of k (the dimensions you reduce to):

$k = \{0, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 650\}$**.**

(e) **Try running the learned model on 4 of the images in the test set and report the results. Give both the binarized input, the true grayscale, and the output of your model.** You may use the code from previous part to visualize the images.

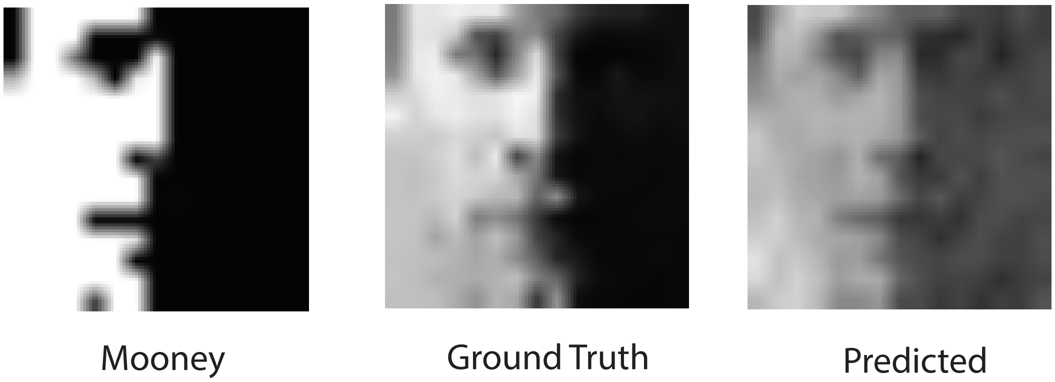


Figure 2: Example results with the input on the left and output on the right

4 Bias-Variance for Ridge Regression (24 points)

Consider the scalar data-generation model:

$$Y = xw^* + Z$$

where x denotes the scalar input feature, Y denotes the scalar noisy measurement, $Z \sim \mathcal{N}(0, 1)$ is standard unit-variance zero-mean Gaussian noise, and w^* denotes the true generating parameter that we would like to estimate.

We are given a set of n training samples $\{x_i, y_i\}_{i=1}^n$ that are generated by the above model with i.i.d. Z_i and distinct x_i. Our goal is to fit a linear model and get an estimate $\widehat{w}$ for the true parameter w^*. For all parts, assume that x_i's are given and fixed (not random).

For a given training set $\{x_i, y_i\}_{i=1}^n$, the ridge-regression estimate for w^* is defined by

$$\widehat{w}_\lambda = \arg\min_{w \in \mathbb{R}} \lambda w^2 + \sum_{i=1}^{n} (y_i - x_i w)^2 \qquad \text{with } \lambda \geq 0.$$

For the rest of the problem, assume that this has been solved and written in the form:

$$\widehat{w}_\lambda = \frac{S_{xy}}{s_x^2 + \lambda} \tag{2}$$

where $S_{xy} = \sum_{i=1}^n x_i Y_i$ and $s_x^2 = \sum_{i=1}^n x_i^2$.

(This is given, no need to rederive it).

(a) (8 pts) **Compute the squared-bias of the ridge estimate $\widehat{w}_\lambda$ defined as follows**

$$\text{Bias}^2(\widehat{w}_\lambda) = (\mathbb{E}[\widehat{w}_\lambda] - w^*)^2. \tag{3}$$

It is fine if your answer depends on w^* or s_x, but it should not depend directly or indirectly on the realizations of the random Z noise. (So, no S_{xy} allowed.)

Hint: First compute the expectation of the estimate $\widehat{w}_\lambda$ over the noises Z in the observation.

(b) (8 pts) **Compute the variance of the estimate $\widehat{w}_\lambda$ which is defined as**

$$\mathrm{Var}(\widehat{w}_\lambda) = \mathbb{E}[(\widehat{w}_\lambda - \mathbb{E}[\widehat{w}_\lambda])^2]. \tag{4}$$

Hint: It might be useful to write $\widehat{w}_\lambda = \mathbb{E}[\widehat{w}_\lambda] + R$ for some random variable R.

(c) (8 pts) **Describe how the squared-bias and variance of the estimate $\widehat{w}_\lambda$ change as we change the value of λ? What happens as $\lambda \to 0$? $\lambda \to \infty$? Is the bias increasing or decreasing? Is the variance increasing or decreasing? In what sense is there a bias/variance tradeoff?**

5 Hospital (25 points)

You work at hospital A. Your hospital has collected patient data to build a model to predict who is likely to get sepsis (a bad outcome). Specifically, the data set contains the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, and associated real number labels $\mathbf{y} \in \mathbb{R}^n$, where n is the number of patients you are learning from and d is the number of features describing each patient. You plan to fit a linear regression model $\widehat{y} = \mathbf{w}^\top \mathbf{x}$ that will enable you to predict a label for future, unseen patients (using their feature vectors).

However, your hospital has only started collecting data a short time ago. Consequently the model you fit is not likely to be particularly accurate. Hospital B has exactly the same set up as your hospital (i.e., their patients are drawn from the same distribution as yours and they have the same measurement tools). For privacy reasons, Hospital B will not share their data. However, they tell you that they have trained a linear model on their own sepsis-relevant data: ($\mathbf{X}_B$ and $\mathbf{y}_B$) and are willing to share their learned model $\widehat{y} = \widehat{\mathbf{w}}_B^\top \mathbf{x}$ with you. In particular, Hospital B shares their entire Gaussian posterior distribution on $\mathbf{w}$ with you: $\mathcal{N}(\widehat{\mathbf{w}}_B, \mathbf{\Psi})$.

(a) (10 pts) Assume that we use the posterior from Hospital B as our own prior distribution for $\mathbf{w} \sim \mathcal{N}(\widehat{\mathbf{w}}_B, \mathbf{\Psi})$. Suppose that our Hospital A model is given by $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where the noise, $\boldsymbol{\epsilon}$, has an assumed distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. **Derive the MAP estimate $\widehat{\mathbf{w}}$ for $\mathbf{w}$ using Hospital A's data $\mathbf{X}, \mathbf{y}$ and the prior information from Hospital B.**

HINT: Recall that traditional ridge regression could be derived from a MAP perspective, where the parameter $\mathbf{w}$ has a zero mean Gaussian prior distribution with a scaled identity covariance. How could you use reparameterization (i.e. change of variables) for the problem here?

(b) (15 pts) Now, for simplicity, consider $d = 1$ so that the w is a scalar parameter. Suppose that instead of giving you their posterior distribution, Hospital B only gave you their mean $\widehat{w}_B$. How can you use this information to help fit your model? **Describe in detail how you should use your own hospital's patient data and combine it with the mean $\widehat{w}_B$ from Hospital B in a procedure to find your own $\widehat{w}$ for predicting sepsis in Hospital A.**

Hint 1: You might want to consider introducing an appropriate hyperparameter and doing what you usually do with hyperparameters.

Hint 2: What does the λ hyperparameter in ridge-regression correspond to from a probabilistic perspective?

6 Ridge regression vs. PCA (24 points)

Assume we are given n training data points $(\mathbf{x}_i, y_i)$. We collect the target values into $\mathbf{y} \in \mathbb{R}^n$, and the inputs into the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ where the rows are the $d-$dimensional feature vectors $\mathbf{x}_i^\top$ corresponding to each training point. Furthermore, assume that $\frac{1}{n}\sum_{i=1}^n \mathbf{x_i} = \mathbf{0}$, $n > d$ and $\mathbf{X}$ has rank d.

In this problem we want to compare two procedures: The first is ridge regression with hyperparameter λ, while the second is applying ordinary least squares after using PCA to reduce the feature dimension from d to k (we give this latter approach the short-hand name k-PCA-OLS where k is the hyperparameter).

Notation: The singular value decomposition of $\mathbf{X}$ reads $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n\times d}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$. We denote by $\mathbf{u}_i$ the n-dimensional column vectors of $\mathbf{U}$ and by $\mathbf{v}_i$ the $d-$dimensional column vectors of $\mathbf{V}$. Furthermore the diagonal entries $\sigma_i = \Sigma_{i,i}$ of $\mathbf{\Sigma}$ satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$. For notational convenience, assume that $\sigma_i = 0$ for $i > d$.

(a) (6 pts) It turns out that the ridge regression optimizer (with $\lambda > 0$) in the $\mathbf{V}$-transformed coordinates

$$\widehat{\mathbf{w}}_{\text{ridge}} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{V}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

has the following expression:

$$\widehat{\mathbf{w}}_{\text{ridge}} = \operatorname{diag}(\frac{\sigma_i}{\lambda + \sigma_i^2})\mathbf{U}^\top\mathbf{y}. \tag{5}$$

Use $\widehat{y}_{test} = \mathbf{x}_{test}^\top \mathbf{V}\widehat{\mathbf{w}}_{\text{ridge}}$ to denote the resulting prediction for a hypothetical $\mathbf{x}_{test}$. Using (5) and the appropriate scalar $\{\beta_i\}$, this can be written as:

$$\widehat{y}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^{d} \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \tag{6}$$

What are the $\beta_i \in \mathbb{R}$ for this to correspond to (5) from ridge regression?

(b) (12 pts) Suppose that we do k-PCA-OLS — i.e. ordinary least squares on the reduced k-dimensional feature space obtained by projecting the raw feature vectors onto the $k < d$ principal components of the covariance matrix $\mathbf{X}^\top \mathbf{X}$. Use $\widehat{y}_{test}$ to denote the resulting prediction for a hypothetical $\mathbf{x}_{test}$,

It turns out that the learned k-PCA-OLS predictor can be written as:

$$\widehat{y}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^{d} \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \tag{7}$$

Give the $\beta_i \in \mathbb{R}$ coefficients for k-PCA-OLS. Show work.

Hint 1: some of these β_i will be zero. Also, if you want to use the compact form of the SVD, feel free to do so if that speeds up your derivation.

Hint 2: some inspiration may be possible by looking at the next part for an implicit clue as to what the answer might be.

(c) (6 pts) For the following part, $d = 5$. The following $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_5)$ (written out to two significant figures) are the results of OLS (i.e. what we would get from ridge regression in the limit $\lambda \to 0$), λ-ridge-regression, and k-PCA-OLS for some $\mathbf{X}, \mathbf{y}$ (identical for each method) and $\lambda = 1, k = 3$. **Write down which procedure was used for each of the three sub-parts below.**

We hope this helps you intuitively see the connection between these three methods.

Hint: It is not necessary to find the singular values of $\mathbf{X}$ *explicitly, or to do any numerical computations at all.*

(i) $\boldsymbol{\beta} = (0.01, 0.1, 0.5, 0.1, 0.01)$

(ii) $\boldsymbol{\beta} = (0.01, 0.1, 1, 0, 0)$

(iii) $\boldsymbol{\beta} = (0.01, 0.1, 1, 10, 100)$

7 Kernel PCA (24 points)

In lectures, discussion, and homework, we learned how to use PCA to do dimensionality reduction by projecting the data to a subspace that captures most of the variability. This works well for data that is roughly Gaussian shaped, but many real-world high dimensional datasets have underlying low-dimensional structure that is not well captured by linear subspaces. However, when we lift the raw data into a higher-dimensional feature space by means of a nonlinear transformation, the underlying low-dimensional structure once again can manifest as an approximate subspace. Linear dimensionality reduction can then proceed. As we have seen in class so far, kernels are an alternate way to deal with these kinds of nonlinear patterns without having to explicitly deal with the augmented feature space. This problem asks you to discover how to apply the "kernel trick" to PCA.

Let $\mathbf{X} \in \mathbb{R}^{n \times \ell}$ be the data matrix, where n is the number of samples and ℓ is the dimension of the raw data. Namely, the data matrix contains the data points $\mathbf{x}_j \in \mathbb{R}^\ell$ as rows

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times \ell}. \tag{8}$$

(a) (5 pts) **Compute $\mathbf{X}\mathbf{X}^\top$ in terms of the singular value decomposition $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{n \times n}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times \ell}$ and $\mathbf{V} \in \mathbb{R}^{\ell \times \ell}$.** Notice that $\mathbf{X}\mathbf{X}^\top$ is the matrix of pairwise Euclidean inner products for the data points. **How would you get $\mathbf{U}$ if you only had access to $\mathbf{X}\mathbf{X}^\top$?**

(b) (7 pts) Given a new test point $\mathbf{x}_{test} \in \mathbb{R}^{\ell}$, one central use of PCA is to compute the projection of $\mathbf{x}_{test}$ onto the subspace spanned by the k top singular vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$.

Express the scalar projection $z_j = \mathbf{v}_j^{\top}\mathbf{x}_{test}$ onto the j-th principal component as a function of the inner products

$$\mathbf{X}\mathbf{x}_{test} = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{x}_{test} \rangle \\ \vdots \\ \langle \mathbf{x}_n, \mathbf{x}_{test} \rangle \end{pmatrix}. \tag{9}$$

Assume that all diagonal entries of $\mathbf{\Sigma}$ are nonzero and non-increasing, that is $\sigma_1 \geq \sigma_2 \geq \cdots > 0$.

Hint: Express $\mathbf{V}^{\top}$ in terms of the singular values $\mathbf{\Sigma}$, the left singular vectors $\mathbf{U}$ and the data matrix $\mathbf{X}$. If you want to use the compact form of the SVD, feel free to do so.

(c) (12 pts) How would you define kernelized PCA for a general kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ (to replace the Euclidean inner product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$)? For example, the RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\delta^2}\right)$.

Describe this in terms of a procedure which takes as inputs the training data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^\ell$ and the new test point $\mathbf{x}_{test} \in \mathbb{R}^\ell$, and outputs the analog of the previous part's z_j coordinate in the kernelized PCA setting. You should include how to compute U from the data, as well as how to compute the analog of $\mathbf{X}\mathbf{x}_{test}$ from the previous part.

Invoking the SVD or computing eigenvalues/eigenvectors is fine in your procedure, as long as it is clear what matrix is having its SVD or eigenvalues/eigenvectors computed. The kernel $k(\cdot, \cdot)$ can be used as a black-box function in your procedure as long as it is clear what arguments it is being given.

8 Multiple Choice Questions (14 points)

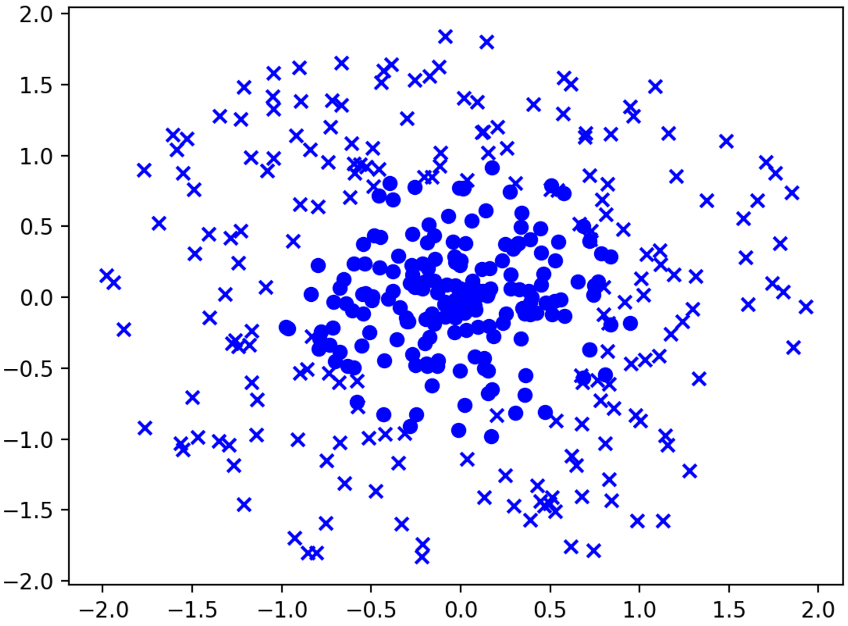
For these questions, select **all** the answers which are correct. You will get full credit for selecting all the right answers. On some questions, real-valued partial credit will be assigned. You will be graded on your **best seven of nine, so feel free to skip up to two of them.**

(a) Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $n \geq d$. Suppose $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the singular value decomposition of $\mathbf{X}$ where $\sigma_i = \Sigma_{i,i}$ are the diagonal entries of $\boldsymbol{\Sigma}$ and satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ while $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$ respectively. **Which of the following is the rank k approximation to $\mathbf{X}$ that is best in the Froebenius norm.** That is, which low rank approximation, $\mathbf{X}_k$, for $\mathbf{X}$ yields the lowest value for $||\mathbf{X} - \mathbf{X}_k||_F^2$?

○ $\sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ ○ $\sum_{i=d-k+1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ ○ $\sum_{i=1}^{k} \sigma_i \mathbf{u}_{n-i} \mathbf{v}_i^\top$

○ $\sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_{n-i}^\top$

(b) Consider a simple dataset of points $(x_i, y_i) \in \mathbb{R}^2$, each associated with a label b_i which is -1 or $+1$. The dataset was generated by sampling data points with label -1 from a disk of radius 1.0 (shown as filled circles in the figure) and data points with label $+1$ from a ring with inner radius 0.8 and outer radius 2.0 (shown as crosses in the figure). **Which set of polynomial features would be best for performing linear regression, assuming at least as much data as shown in the figure?**



○ $1, x_i$

○ $1, x_i, y_i$

○ $1, x_i, y_i, x_i^2, x_i y_i, y_i^2$

○ $1, x_i, y_i, x_i^2, x_i y_i, y_i^2, x_i^3, y_i^3, x_i^2 y_i, x_i y_i^2$

(c) **Which of the following is a valid kernel function** for vectors of the same length, $\mathbf{x}$ and $\mathbf{y}$?

○ $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$

○ $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{2}||\mathbf{x}-\mathbf{y}||_2^2}$

○ $k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^\top \mathbf{y})^p$ for some degree p

○ $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) - k_2(\mathbf{x}, \mathbf{y})$ for valid kernels k_1 and k_2.

(d) During training of your model, both independent variables in the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and dependent target variables $\mathbf{y} \in \mathbb{R}^n$ are corrupted by noise. At test time, the data points you are computing predictions for, $\mathbf{x}_{test}$, are noiseless. **Which method(s) should you use to estimate the value of $\hat{\mathbf{w}}$ from the training data in order to make the most accurate predictions $\mathbf{y}_{test}$ from the noiseless test input data, $\mathbf{X}_{test}$?** Assume that you make predictions using $\mathbf{y}_{test} = \mathbf{X}_{test}\hat{\mathbf{w}}$.

○ OLS

○ Ridge regression

○ Weighted Least Squares

○ TLS

(e) Assume you have n input data points, each with d high quality features ($\mathbf{X} \in \mathbb{R}^{n \times d}$) and associated labels ($\mathbf{y} \in \mathbb{R}^n$). Suppose that $d \gg n$ and that you want to learn a linear predictor. **Which of these approaches would help you to avoid overfitting?**

○ Preprocess $\mathbf{X}$ using $k \ll n$ random projections

○ Preprocess $\mathbf{X}$ using PCA with $k \ll n$ components.

○ Preprocess $\mathbf{X}$ using PCA with n components.

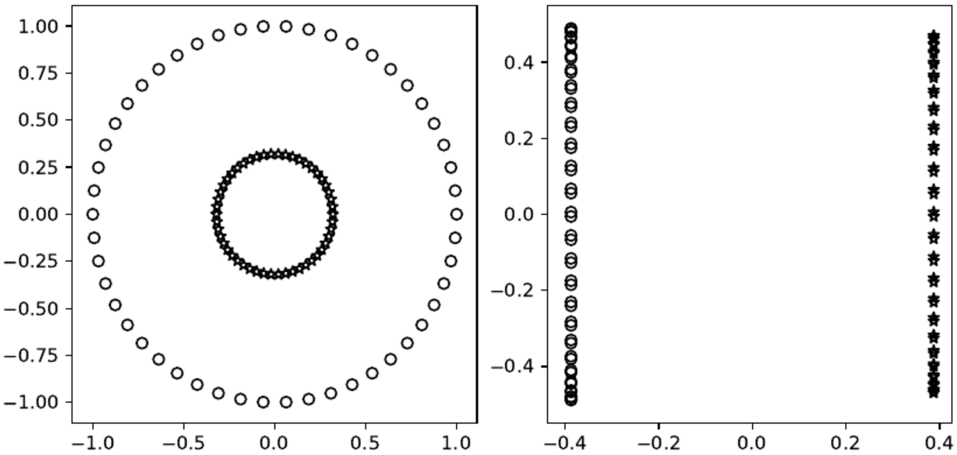
○ Add polynomial features

○ Use a kernel approach

○ Add a ridge penalty to OLS

○ Do weighted least squares

(f) **Which methods could yield a transformation to go from the two-dimensional data on the left to the two-dimensional data on the right?**



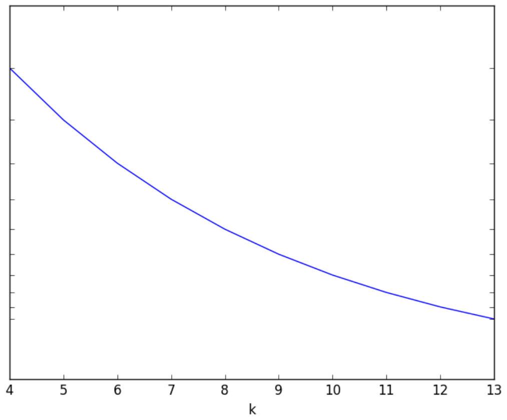
○ Random projections

○ PCA

○ Use of a kernel

○ Adding polynomial features

(g) Your friend is training a machine learning model to predict disease severity based on k different health indicators. She generates the following plot, where the value of k is on the x axis.



Which of these might the y axis represent?

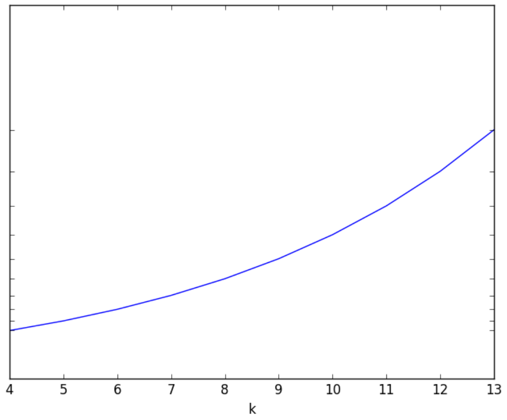
○ Training Error

○ Validation Error

○ Bias

○ Variance

(h) Your friend is training a machine learning model to predict disease severity based on k different health indicators. She generates the following plot, where the value of k is on the x axis.



Which of these might the y axis represent?

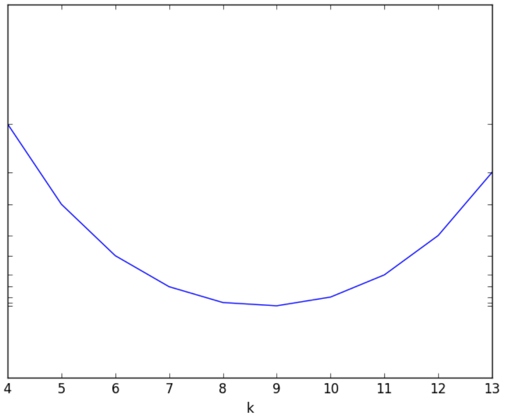
○ Training Error

○ Validation Error

○ Bias

○ Variance

(i) Your friend is training a machine learning model to predict disease severity based on k different health indicators. She generates the following plot, where the value of k is on the x axis.



Which of these might the y axis represent?

○ Training Error

○ Validation Error

○ Bias

○ Variance

9 Your Own Question

Write your own question, and provide a thorough solution.

Writing your own problems is a very important way to really learn the material. The famous "Bloom's Taxonomy" that lists the levels of learning is: Remember, Understand, Apply, Analyze, Evaluate, and Create. Using what you know to create is the top-level. We rarely ask you any HW questions about the lowest level of straight-up remembering, expecting you to be able to do that yourself. (e.g. make yourself flashcards) But we don't want the same to be true about the highest level.

As a practical matter, having some practice at trying to create problems helps you study for exams much better than simply counting on solving existing practice problems. This is because thinking about how to create an interesting problem forces you to really look at the material from the perspective of those who are going to create the exams.

Besides, this is fun. If you want to make a boring problem, go ahead. That is your prerogative. But it is more fun to really engage with the material, discover something interesting, and then come up with a problem that walks others down a journey that lets them share your discovery. You don't have to achieve this every week. But unless you try every week, it probably won't happen ever.