CS 189 Introduction to Machine Learning
Fall 2018 HW2

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 to assignment on Gradescope, "HW2 Write-Up". If there are graphs, include those graphs in the correct sections. Do not simply reference your appendix.

(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 **Wednesday, September 12 at 10 p.m.**.

2 Prediction error of Ridge Regression

(a) Let A be a $d \times n$ matrix and B be a $n \times d$ matrix. For any $\mu > 0$, **show that** $(AB + \mu I)^{-1}A = A(BA + \mu I)^{-1}$**.**, if $AB + \mu I$ and $BA + \mu I$ are invertible.

(b) Let $X \in \mathbb{R}^{n \times d}$ be n samples of d features, and $y \in \mathbb{R}^n$ be the corresponding n samples of the quantity that you would like to predict with regression. Let

$$\widehat{\theta}_\lambda = \arg\min_\theta \|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2,$$

for $\lambda > 0$, be the solution to the ridge regression problem.

Show that $\widehat{\theta}_\lambda = X^\top(XX^\top + \lambda I)^{-1}y$.

(c) Suppose X has the singular value decomposition $U\Sigma V^\top$, where $\Sigma = \mathrm{diag}(s_1, \cdots, s_d)$, $s_i \geq 0$. Using part b), **show that** $\widehat{\theta}_\lambda = VDU^\top y$**, where** D **is a diagonal matrix to be determined.**

(d) Let $\widehat{y}_\lambda = X\widehat{\theta}_\lambda$ be the predictions made by the ridge regressor $\widehat{\theta}_\lambda$. **Show that** $\widehat{y}_\lambda = Py$**, where** P **is a matrix to be determined. Comment on what** P **is doing to** y**.**

(e) Now suppose we have $y = X\theta_* + z$, where $\theta_* \in \mathbb{R}^d$ and $z = \mathcal{N}(0, \sigma^2 I) \in \mathbb{R}^n$ ($\sigma > 0$). Further suppose that X is an orthogonal matrix, that is, $X^\top X = I$.

$\mathbb{E}\|X(\widehat{\theta}_\lambda - \theta_*)\|^2$ is the expected squared difference between the predictions made by the ridge regressor $\widehat{y}_\lambda$ (see previous part) and y, where the expectation is taken with respect to the random variable z. ($\|\cdot\|$ denotes the $\ell 2$ norm.) We can think of this expression as the (in-sample) prediction error.

Show that $\mathbb{E}\|X(\widehat{\theta}_\lambda - \theta_*)\|^2 = \frac{1}{(1+\lambda)^2}\left(\lambda^2\|\theta_*\|^2 + d\sigma^2\right)$**.**

(f) **What is the** λ^* **that you should pick to minimize the prediction error you computed in part e)? Comment on how** d, σ^2**, and** θ_* **each affect the optimal choice of the regularization parameter** λ**.**

(g) **What is the prediction error** $\mathbb{E}\|X(\widehat{\theta}_{\lambda_*} - \theta_*)\|^2$ **using the optimized** λ^***? Compare this prediction error to the prediction error for Ordinary Least Squares**

$$\mathbb{E}\|X(\widehat{\theta}_{OLS} - \theta_*)\|^2 = d\sigma^2,$$

which you computed in problem 5 f) of Homework 1.

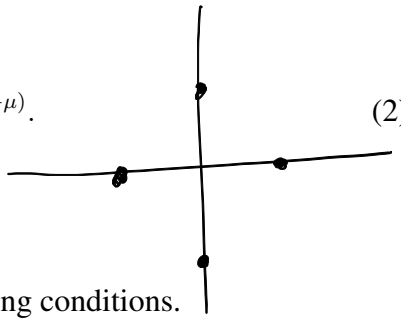
3 Independence and Multivariate Gaussians

As described in lecture, a covariance matrix $\Sigma \in \mathbb{R}^{N \times N}$ for a random variable $X \in \mathbb{R}^N$ with the following values, where $\mathrm{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$ is the covariance between the i-th and j-th elements of the random vector X:

$$\Sigma = \begin{bmatrix} \text{cov}(X_1, X_1) & \dots & \text{cov}(X_1, X_n) \\ \dots & & \dots \\ \text{cov}(X_n, X_1) & \dots & \text{cov}(X_n, X_n) \end{bmatrix}. \tag{1}$$

Recall that the density of an N dimensional Multivariate Gaussian Distribution $\mathcal{N}(\mu, \Sigma)$ is defined as follows when Σ is positive definite:

$$f(x) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}. \tag{2}$$



Here, $|\Sigma|$ denotes the determinant of the matrix Σ.

(a) Consider the random variables X and Y in $\mathbb{R}$ with the following conditions.

(i) X and Y can take values $\{-1, 0, 1\}$.

(ii) When X is 0, Y takes values 1 and -1 with equal probability ($\frac{1}{2}$). When Y is 0, X takes values 1 and -1 with equal probability ($\frac{1}{2}$).

(iii) Either X is 0 with probability ($\frac{1}{2}$), or Y is 0 with probability ($\frac{1}{2}$).

Are X and Y uncorrelated? Are X and Y independent? Prove your assertions. *Hint:* Write down the joint probability of (X, Y) for each possible pair of values they can take.

(b) For $X = [X_1, \cdots, X_n]^\top \sim \mathcal{N}(\mu, \Sigma)$, **verify that if X_i, X_j are independent (for all $i \neq j$), then Σ must be diagonal, that is, X_i, X_j are uncorrelated.**

(c) Let $N = 2$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$. Suppose $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$. **Show that X_1, X_2 are independent if $\beta = 0$.** Recall that two continuous random variables W, Y with joint density $f_{W,Y}$ and marginal densities f_W, f_Y are independent if $f_{W,Y}(w, y) = f_W(w) f_Y(y)$.

(d) Consider a data point x drawn from a N-dimensional zero mean Multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, as shown above. Assume that Σ^{-1} exists. **Prove that there exists matrix $A \in R^{N,N}$ such that $x^\top \Sigma^{-1} x = \|Ax\|_2^2$ for all vectors x. What is the matrix A?**

(e) Let's constrain x to be on the unit sphere. In other words, the ℓ_2 norm (or magnitude) of vector x is 1 ($\|x\|_2 = 1$). In this case, **what are the maximum and minimum values of $\|Ax\|_2^2$? In other words, $\max_{x:\|x\|_2=1} \|Ax\|_2^2$ and $\min_{x:\|x\|_2=1} \|Ax\|_2^2$?**

(f) If we had $X_i \perp\!\!\!\perp X_j \ \forall i, j$ ($\perp\!\!\!\perp$ denotes independence), **what is the intuitive meaning for the maximum and minimum values of $\|Ax\|_2^2$?** Suppose you wanted to choose an x on the unit sphere to maximize the density function $f(x)$ in Eq (2); **what x should you choose?**

4 Blair and their giant peaches

Make sure to submit the code you write in this problem to "HW2 Code" on Gradescope.

Blair is a mage testing how long they can fly a collection of giant peaches. They has n training peaches – with masses given by $x_1, x_2, \ldots x_n$ – and flies these peaches once to collect training data. The experimental flight time of peach i is given by y_i. They believes that the flight time is well approximated by a polynomial function of the mass

$$y_i \approx w_0 + w_1 x_i + w_2 x_i^2 \cdots + w_D x_i^D$$

where their goal is to fit a polynomial of degree D to this data. Include all text responses and plots in your write-up.

(a) **Show how Blair's problem can be formulated as a linear regression problem.**

(b) You are given data of the masses $\{x_i\}_{i=1}^n$ and flying times $\{y_i\}_{i=1}^n$ in the "x_train" and "y_train" keys of the file `1D_poly.mat` with the masses centered and normalized to lie in the range $[-1, 1]$. **Write a script by completing `b.py` to do a least-squares fit (taking care to include a constant term) of a polynomial function of degree D to the data.** Letting f_D denote the fitted polynomial, **plot the average training error** $R(D) = \frac{1}{n}\sum_{i=1}^n (y_i - f_D(x_i))^2$ **against D in the range $D \in \{0, 1, 2, 3, \ldots, n-1\}$.** You may not use any library other than `numpy` and `numpy.linalg` for computation.

(c) **How does the average training error behave as a function of D, and why? What happens if you try to fit a polynomial of degree n with a standard matrix inversion method?**

(d) Blair has taken CS189 so decides that they needs to run another experiment before deciding that their prediction is true. They runs another fresh experiment of flight times using the same peaches, to obtain the data with key "y_fresh" in 1D_POLY.MAT. Denoting the fresh flight time of peach i by $\tilde{y}_i$, **by completing `c.py`, plot the average error** $\tilde{R}(D) = \frac{1}{n}\sum_{i=1}^n (\tilde{y}_i - f_D(x_i))^2$ **for the same values of D as in part (b) using the polynomial approximations f_D also from the previous part. How does this plot differ from the plot in (b) and why?**

(e) **How do you propose using the two plots from parts (b) and (d) to "select" the right polynomial model for Blair?**

(f) Blair has a new hypothesis – the flying time is actually a function of the mass, smoothness, size, and sweetness of the peach, and some multivariate polynomial function of all of these parameters. A D-multivariate polynomial function looks like

$$f_D(\mathbf{x}) = \sum_j \alpha_j \prod_i x_i^{p_{ji}},$$

where $\forall j : \sum_i p_{ji} \leq D$. Here α_j is the scale constant for jth term and p_{ji} is the exponent of x_i in jth term. The data in `polynomial_regression_samples.mat` (100000×5) with columns corresponding to the 5 attributes of the peach. **Use 4-fold cross-validation to decide**

which of $D \in \{0, 1, 2, 3, 4, 5, 6\}$ is the best fit for the data provided. For this part, compute the polynomial coefficients via ridge regression with penalty $\lambda = 0.1$, instead of ordinary least squares. You are not allowed to use any library other than `numpy` and `numpy.linalg`. Write your implementation by completing `fg.py`.

(g) Now **redo the previous part, but use 4-fold cross-validation on all combinations of $D \in \{1, 2, 3, 4, 5, 6\}$ and $\lambda \in \{0.05, 0.1, 0.15, 0.2\}$** - this is referred to as a grid search. **Find the best D and λ that best explains the data using ridge regression. Print the average training/validation error per sample for all D and λ.** Again, write your implementation by completing `fg.py`.