CS 189 Introduction to Machine Learning
Fall 2017 HW2

This homework is due **Friday, September 8 at 10 p.m.**.

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW[n] Write-Up"
2. Submit all code needed to reproduce your results, "HW[n] Code".
3. Submit your test set evaluation results, "HW[n] Test Set".

After you've submitted your homework, be sure to watch out for the self-grade form.

(a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. 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:

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.

2 Geometry of Ridge Regression

One way to interpret "ridge regression" is as the Lagrangian form of a constrained problem.

(a) Given a matrix $X \in \mathbb{R}^{n \times d}$ and a vector $\vec{y} \in \mathbb{R}^n$, define the optimization problem

$$\begin{aligned} &\text{minimize } \|\vec{y} - X\vec{w}\|_2^2. \\ &\text{subject to } \|\vec{w}\|_2^2 \leq \beta^2. \end{aligned} \tag{1}$$

Using the spirit of the method of Lagrange multipliers (wherein we replace a constraint with a penalty on the thing that we are constraining, and then adjust the level of the penalty until the solution ends up satisfying the constraint. These penalties are sometimes referred to as "shadow prices," especially in the economics literature.), rewrite the constrained optimization problem an an unconstrained optimization with the constraint incorporated within the objective function.

Recall that ridge regression is given by the unconstrained optimization problem

$$w = \arg\min_w \|\vec{y} - X\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2. \tag{2}$$

Hence, by performing ridge regression with penalty λ, we are essentially solving a constrained least squares problem with our parameter having bounded Euclidean norm β. **Qualitatively, how would increasing β be reflected in the desired penalty λ of ridge regression?**

(b) One reason why we might want to have small weights $\vec{w}$ has to do with the sensitivity of the predictor to its input. Let $\vec{x}$ be a d-dimensional list of features corresponding to a new test point. Our predictor is $\vec{w}^\top \vec{x}$. **By how much can our prediction change if nature added an arbitrary disturbance vector of length ε to the test point's features $\vec{x}$?**

(c) **Derive that the solution to ridge regression** (2) **is given by $\hat{w}_r = (X^\top X + \lambda I)^{-1} X^\top y$. What happens when $\lambda \to \infty$?** It is for this reason that sometimes regularization is referred to as "shrinkage."

(d) Note that in computing $\hat{w}_r$, we are trying to invert the matrix $X^\top X + \lambda I$ instead of the matrix $X^\top X$. **If $X^\top X$ has eigenvalues $\lambda_1, \ldots, \lambda_d$, what are the eigenvalues of $X^\top X + \lambda I$? Comment on why adding the regularizer term λI can improve the inversion operation numerically.**

(e) Let $d = 3$, $n = 5$, and let the eigenvalues of $X^\top X$ be given by 1000, 1 and 0.001. We must now choose between two regularization parameters $\lambda_1 = 100$ and $\lambda_2 = 0.5$. **Which do you think is a better choice for this problem and why?**

(f) Another advantage of ridge regression can be seen for under-determined systems. Say we have the data drawn from a 5 parameter model, but only have 4 samples of it, i.e. $X \in \mathbb{R}^{4 \times 5}$. Now this is clearly an underdetermined system, since $n < d$. **Show that ridge regression with $\lambda > 0$ results in a unique solution, whereas ordinary least squares has an infinite number of solutions.**

Hint: To make this point, it may be helpful to expand any vector w as $w = w_0 + X^\top a$ for $w_0 \in \mathsf{nullspace}(X)$ and some a.

(g) (**BONUS**) For the previous part, **what will the answer be if you take the limit $\lambda \to 0$ for ridge regression?**

(h) Tikhonov regularization is a general term for ridge regression, where the constraint set takes the form of an ellipsoid instead of a ball. In other words, we solve the optimization problem

$$w = \arg\min_w \|y - Xw\|_2^2 + \lambda \|\Gamma w\|_2^2$$

for some full rank matrix $\Gamma \in \mathbb{R}^{d \times d}$. **Derive a closed form solution to this problem.**

3 Polynomials and invertibility

Consider using a D-degree polynomial to fit a function $y = f(x)$ with n training samples, where both x and y are scalar. We know that this is equivalent to performing linear regression with a feature matrix F constructed from the n training data sampling positions $x_1, \ldots, x_n$. Assume all the training sampling positions are non-zero, and let this mapping be given by $F = [\vec{p}_D(x_1), \ldots, \vec{p}_D(x_n)]^T$ where $\vec{p}_D(x) = [x^0, x^1, \ldots, x^D]^T$.

(a) For $n = 2$ and $D = 1$, **show that the matrix F has full rank iff $x_1 \neq x_2$.**

(b) More generally, let us now show that the columns of F are linearly independent provided the sampling data points are distinct and $n \geq D+1$. It suffices to consider the case $n = D+1$ and so assume that from this point forward for all the questions about univariate polynomials.

We do this by performing the following operations in sequence. From the matrix F, subtract the first row from rows 2 through n to obtain matrix F'.

Is it true that $\det(F) = \det(F')$**?**

Hint: Think about representing the row subtraction operation using a matrix multiplication, and argue why this additional matrix must have determinant 1. (What are the eigenvalues of a triangular matrix?)

(c) Perform the following sequence of operations to F', and obtain the matrix F''.

i) Subtract $x_1 * \mathsf{column}_{n-1}$ from column_n.

ii) Subtract $x_1 * \mathsf{column}_{n-2}$ from column_{n-1}.

$\vdots$

n-1) Subtract $x_1 * \mathsf{column}_1$ from column_2.

Write out the matrix F'' and argue why $\det(F') = \det(F'')$**.**

(d) For any square matrix $A \in \mathbb{R}^{d \times d}$ and a matrix

$$B = \begin{bmatrix} 1 & \vec{0}^\top \\ \vec{0} & A \end{bmatrix},$$

argue that the $d+1$ eigenvalues of B are given by $\{1, \lambda_1(A), \lambda_2(A), \ldots, \lambda_d(A)\}$, and conclude that $\det(B) = \det(A)$**.** Here, $\vec{0}$ represents a column vector of zeros in $\mathbb{R}^d$.

(e) **Use the above parts to show by induction that we have** $\det(F) = \prod_{1 \le i < j \le n}(x_j - x_i)$**.** Consequently, the matrix X is full rank unless two data points are equal.

Hint: First show that

$$\det(F) = \left(\prod_{i=2}^{n}(x_i - x_1)\right)\det([\vec{p}_{D-1}(x_2), \vec{p}_{D-1}(x_3), \ldots, \vec{p}_{D-1}(x_n)]^T).$$

Hint Hint: You can use the fact that multiplying a row of a matrix by a constant scales the determinant by this constant. (A fact that is clear from the oriented volume interpretation of determinants.)

(f) Let us now extend this argument to features from a multidimensional space of dimension ℓ, using a multivariate polynomial of degree D.

Using a stars and bars (link is here if you did not take CS70) argument, **show that now, we have** $\binom{D+\ell}{\ell}$ **features for each sampling point.**

(g) Choose n sample points $\{\vec{x}_i\}_{i=1}^n$ with $\vec{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,\ell})^T$, and stack up all their features like in part (a) to form the feature matrix F_ℓ.

First, we will show that for some choice of distinct sampling points, this may not be full rank. Let us now form a particular instance of the matrix F_ℓ by choosing $x_{i,1} = x_{i,2} = \cdots = x_{i,\ell} = \alpha_i$ for distinct α_i as $i \in \{1, 2, 3 \ldots, n\}$. **Show that this leads to an** F_ℓ **with linearly dependent columns no matter how many samples** n **you take.**

(h) To show that the matrix can be full rank, **pick a set of sampling points** $\vec{x}_i$ **to show that there is a way to choose the samples to get the matrix** F_ℓ **to be full column rank.** You are allowed to pick as many sample points as you want. Although we are only asking you to show that there exists a way to sample to achieve full rank with enough samples taken, it turns out to be true that the F_ℓ matrix will be full rank as long as $n = \binom{D+\ell}{\ell}$ "generic" points are chosen.

Hint: Leverage earlier parts of this problem if you can.

4 Polynomials and approximation

For a p-times differentiable function $f : \mathbb{R} \to \mathbb{R}$, the Taylor series expansion of order $m \le p-1$ about the point x_0 is given by

$$f(x) = \sum_{i=0}^{m} \frac{1}{i!} f^{(i)}(x_0)(x - x_0)^i + \frac{1}{(m+1)!} f^{(m+1)}(a(x))(x - x_0)^{m+1}. \tag{3}$$

Here, $f^{(m)}$ denotes the mth derivative of the function f, and $a(x)$ is some value between x_0 and x.

The last term of this expansion is tyically referred to as the approximation-error term when approximating $f(x)$ by an m-th degree polynomial. For functions f whose derivatives are bounded in the neighborhood of interest, if we have $|f^{(m)}(x)| \le T$ for $x \in (x_0 - s, x_0 + s)$, then we know that for $x \in (x_0 - s, x_0 + s)$ that $|f(x) - \sum_{i=0}^{m} \frac{1}{i!} f^{(i)}(x_0)(x - x_0)^i| \le \frac{T(x-x_0)^{m+1}}{(m+1)!}$.

(a) **Compute the 2nd, 3rd, and 4th order Taylor approximation of the following functions about the point $x_0 = 0$. Also bound the error of approximation at $x = 3$.**

i) e^x

ii) $\sin x$

(b) Let us say we would like an accurate polynomial approximation of the functions in part (a) for all $x \in [-3, 3]$. In other words, we want a polynomial of degree D (call it ϕ_D) such that $|f(x) - \phi_D(x)| \leq \varepsilon$ for all $x \in [0, 3]$. **How large does D need to be as a function of ε for such a guarantee for the two choices of $f(x)$ in part (a)?**

(c) **What is $\lim_{m\to\infty} \frac{(x-x_0)^{m+1}}{(m+1)!}$?**

Conclude that a univariate polynomial of high enough degree can approximate any function that is sufficiently smooth. This is the power of using polynomial features, even when we don't know the underlying function f that is generating our data! This universal approximation property gives us some justification for using polynomial features. (Later, we will see that neural networks are also universal function approximators.)

(d) Now let's extend this idea of approximating functions with polynomials to multivariable functions. The Taylor series expansion for a function $f(x, y)$ about the point (x_0, y_0) is given by

$$\begin{aligned} f(x,y) = f(x_0,y_0) + f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0) + \\ \frac{1}{2!}[f_{xx}(x_0,y_0)(x-x_0)^2 + f_{xy}(x_0,y_0)(x-x_0)(y-y_0) + \\ f_{yx}(x_0,y_0)(x-x_0)(y-y_0) + f_{yy}(x_0,y_0)(y-y_0)^2] + \ldots \end{aligned} \tag{4}$$

where $f_x = \frac{\partial f}{\partial x}$, $f_y = \frac{\partial f}{\partial y}$, $f_{xx} = \frac{\partial^2 f}{\partial x^2}$, $f_{yy} = \frac{\partial^2 f}{\partial y^2}$, and $f_{xy} = \frac{\partial^2 f}{\partial x \partial y}$

As you can see, the Taylor series for multivariate functions quickly becomes unwieldy after the second order. Let's try to make the series a little bit more manageable. **Using matrix notation, write the expansion for a function of two variables in a more compact form up to the second order terms where $f(\vec{x}) = f(x, y)$ with $\vec{x} = [x, y]^T$ and $\vec{x}_0 = [x_0, y_0]$. Clearly define any additional vectors and matrices that you use.**

(e) To see that we can do universal approximation, first just consider that we want to approximate the function in a single straight line path. For this problem, we will use the function

$f(\vec{x}) = e^x \sin y$

where $\vec{x} = [x, y]^T$. We're interested in the path $\vec{x}(t) = [\frac{1}{\sqrt{2}}t, \frac{1}{\sqrt{2}}t]^T$. **Write the 3rd order Taylor expansion of $f(\vec{x}(t))$ for the variable t about the point $t_0 = 0$.** Use the chain rule.

(f) Similar to part **(b)**, **determine the degree D for the polynomial $\phi_D(\vec{x}(t))$ from the Taylor series about the point $t_0 = 0$ such that $|f(\vec{x}(t)) - \phi_D(\vec{x}(t))| \leq \varepsilon$ on the interval $t \in [0, 3]$ for the function from the previous part.**

(g) BONUS: **Sketch how the argument above can be extended to show that multivariate polynomials are universal function approximators for sufficiently smooth functions of many variables.**

5 Jaina and her giant peaches

NOTE: In response to student questions, a new data file 1D_plot_new.mat was added to the website at 3.30PM on September 7th. You will be awarded full credit for implementations on the first data file 1D_poly.mat, but the second dataset 1D_poly_new.mat is bonus. Please use the new file to see the expected behavior of errors and for extra credit!

In another alternative universe, Jaina is a mage testing how long she can fly a collection of giant peaches. She has n training peaches – with masses given by $x_1, x_2, \ldots x_n$ – and flies all the peaches once to collect training data. The experimental flight time of peach i is given by y_i. She believes that the flight time is well approximated by a polynomial function of the mass, and her 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 Jaina'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, respectively (the new data is in the file 1D_POLY_NEW.MAT), with the masses centered and normalized to lie in the range $[-1,1]$. **Write a routine 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 $D \in \{1,2,3,\ldots,n-1\}$.**

(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) Jaina has taken Mystical Learning 189, and so decides that she needs to run another experiment before deciding that her prediction is true. She 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 $\tilde{y}_i$, **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 Jaina?**

(f) Jaina 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. 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 $D \in \{1,2,3,4\}$ 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.

(g) Now **redo the previous part, but use 4-fold cross-validation on all combinations of $D \in \{1,2,3,4\}$ 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.**

6 Your Own Question

Write your own question, and provide a thorough solution.

This problem should show your understanding of a key concept in the class.

Writing your own problems is a very important way to really learn 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.