CSC 580 Homework 2

Due: 10/15 (Saturday) 5:00pm

Instructions:

- **(NEW)**: If you use math symbols, please define it clearly before you use it (unless they are standard from the lecture).
- **(NEW)**: You must provide the full source code for whatever problem you use programming. Please email them to csc580homeworks@gmail.com.
- Every single subproblem will be worth 10 points.
- Submit your homework on time to gradescope. NO LATE DAYS, NO LATE SUBMISSIONS ACCEPTED.
- The submission must be one single PDF file (use Acrobat Pro from the UA software library if you need to merge multiple PDFs).
- Copy and paste the code into the pdf submission (you can use Microsoft Word and use fixed-width fonts).
  - You can use word processing software like Microsoft Word or LaTeX.
  - You can also hand-write your answers and then scan it. If you use your phone camera, I recommend using TurboScan (smartphone app) or similar ones to avoid looking slanted or showing the background.
  - Watch the video and follow the instruction: https://youtu.be/KMPoby5g_nE .
- Collaboration policy: do not discuss answers with your classmates. You can discuss HW for the clarification or any math/programming issues at a high-level. If you do get help from someone, please write it down in the answer.
- Please use the problem & subproblem numbering of this document; do not recreate or renumber them.

Problem 1. Math

(a) Prove that

(1) the max uncertainty $F_1(p) = 1 - \max_{i=1}^{k} p_i$,
(2) the Gini index-based uncertainty $F_2(p) = 1 - \sum_{i=1}^{k} p_i^2$,
(3) and the entropy uncertainty $F_3(p) = \sum_{i=1}^{k} p_i \log(1/p_i)$,

are all concave functions. Clearly describe which properties of convexity/concavity you are using.

(b) Consider the standard regression setting with the squared loss. Let us assume that the conditional distribution $Y|X = x$ follows the Gaussian distribution with variance σ^2 at some mean that is a function of x. You can use any symbol you like to denote this function in your solution, such as f or g. Describe the regression function (i.e., the Bayes optimal function) in this case (the answer will not involve $\min$ or $\arg\min$ but rather an expectation). Compute the Bayes risk of the regression function. **Note**: just writing the definitions will result in 0 point. **Hint**: the Bayes risk will involve σ^2 in some ways.

Problem 2. Linear Regression

The Wisconsin State Climatology Office keeps a record on the number of days Lake Mendota was covered by ice at `http://www.aos.wisc.edu/~sco/lakes/Mendota-ice.html`. Same for Lake Monona: `http://www.aos.wisc.edu/~sco/lakes/Monona-ice.html`. As with any real problems, the data is not as clean or as organized as one would like for machine learning. Curate two clean data sets for each lake, respectively, starting from 1855-56 and ending in 2018-19. Let x be the year: for 1855-56, $x = 1855$; for 2017-18, $x = 2017$; and so on. Let y be the ice days in that year: for Mendota and 1855-56, $y = 118$; for 2017-18, $y = 94$; and so on. Some years have multiple freeze thaw cycles such as 2001-02, that one should be $x = 2001, y = 21$.

(a) Plot year vs. ice days for the two lakes as two curves in the same plot. Produce another plot for year vs. $y_{Monona} - y_{Mendota}$.

(b) Split the datasets: $x \le 1970$ as training, and $x > 1970$ as test. (Comment: due to the temporal nature this is NOT an iid split. But we will work with it.) On the training set, compute the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and the sample standard deviation $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ for the two lakes, respectively.

(c) Using training sets, train a linear regression model

$$\hat{y}_{Mendota} = \beta_0 + \beta_1 x + \beta_2 y_{Monona}$$

to predict $y_{Mendota}$. Note: we are treating y_{Monona} as an observed feature. In other words, each example is represented by feature $z = (1, x, y_{Monona})$ and label $y_{Mendota}$, and the regression model is:

$$\hat{y}_{Mendota} = \langle \beta, z \rangle$$

for $\beta = (\beta_0, \beta_1, \beta_2)^\top$. Do this by finding the closed-form ordinary least squares (OLS) solution for $\beta = (\beta_0, \beta_1, \beta_2)^\top$ (no regularization):

$$\min_\beta \frac{1}{n} \sum_{i=1}^{n} (z_i^\top \beta - y_i)^2.$$

Give the OLS formula in matrix form (define your matrices), then give the OLS value of $\beta_0, \beta_1, \beta_2$.

(d) Using the MLE above, give the mean squared error on the test set; i.e., $\frac{1}{m} \sum_{i=1}^{m} (z_i^\top \beta - y_i)^2$ where $\{(z_i, y_i)\}_{i=1}^m$ is the test set. (You will need to use the Monona test data as observed features.)

(e) "Reset" to (c), but this time use gradient descent to learn the β's. Recall our objective function is the mean squared error on the training set:

$$F(\beta) := \frac{1}{n} \sum_{i=1}^{n} (z_i^\top \beta - y_i)^2.$$

Derive the gradient of F with respect to β.

(f) Implement gradient descent. Initialize $\beta_0 = \beta_1 = \beta_2 = 0$. Use a fixed stepsize parameter $\eta = 0.1$ and print the first 10 iteration's objective function value. Use the stopping criterion learned in the class with tolerance 10^{-4}; report the final β and its training objective function value. Compare the β's to the closed-form OLS solution. Try smaller η values and tell us what happens.
Hint: Update $\beta_0, \beta_1, \beta_2$ simultaneously in an iteration. Don't use a new β_0 to calculate β_1, and so on.

(g) As preprocessing, normalize your year and Monona features (but not $y_{Mendota}$). Here, normalization means the standardization where you transform the features so that they are mean 0 and variance 1. Then repeat (f).

(h) "Reset" to (c) (no normalization, use closed-form solution), but train a linear regression model without using Monona:

$$\hat{y}_{Mendota} = \gamma_0 + \gamma_1 x.$$

i. Interpret the sign of γ_1.

ii. Some analysts claim that because the closed-form solution β_1 in (c) is positive, fixing all other factors, as the years go by the number of Mendota ice days will increase, namely the model in (c) indicates a cooling trend. Discuss this viewpoint, relate it to question (h) i.

Problem 3. Ridge regression.

(a) Derive the closed-form solution in matrix form for the ridge regression problem:

$$\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} (z_i^\top \beta - y_i)^2 \right) + \lambda \|\beta\|_A^2$$

where $\beta = (\beta_0, \beta_1, \beta_2)^\top$,

$$\|\beta\|_A^2 := \beta^\top A \beta$$

and

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

This A matrix has the effect of NOT regularizing the bias β_0, which is standard practice in ridge regression. Note: Derive the closed-form solution with step by step explanations, do not blindly copy lecture notes. Writing your solutions in terms of z_i's would be fine. If you like, you can also represent each z_i as $z_i = (1, x_i)$ in this problem where each $x_i \in \mathbb{R}^2$, and write your solutions in terms of the x_i's.)

(b) Let $\lambda = 1$ and tell us the value of β from your ridge regression model when you use the data from Problem 2(c).

Problem 4. Paired t-test

You lead a team in a startup company that provides a search engine for digital document search in libraries. You have come up with a new ranking algorithm B for the search results. You believe that B improves upon the existing algorithm A. You now want to test if the users actually like the result from B in a statistically meaningful way. You have crowdsourced 12 evaluators, showed them the search results from A and B side by side, and asked them to provide rating scores from 1 to 5. Here is the data:

A	B
1	1
2	4
1	3
2	1
4	5
3	5
2	4
3	4
4	2
1	3
2	4
1	2

(a) For each algorithm, construct a 95% confidence interval for the mean rating under the assumption that each trial follows an i.i.d. Gaussian distribution. Here, one trial is one rating from an evaluator. Use python code and report both the code and the result. Please state whether the two confidence intervals overlap or not. Based on your answer, can you claim that one algorithm is better than the other in a statistically meaningful way?

(b) You are asked to test the null hypothesis that the two algorithms have no difference in user ratings. Perform the two-sided paired t-test and tell us if you were able to reject the null hypothesis with the significance level $\alpha = 0.05$ (i.e., with 95% confidence). Based on the result, discuss the potential benefit of using the paired t-test as opposed to constructing individual confidence intervals as in (a).

Problem 5. SVM and kernels.

(a) Derive the dual problem of the *homogeneous* soft-margin SVM:

$$\begin{aligned} &\min_{w} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{such that } &\; y_i w^\top x_i \geq 1 - \xi_i, \forall i \\ &\; \xi_i \geq 0, \forall i \end{aligned}$$

Important: Please provide step-by-step justifications in your own words.

Example i is said to be a *support vector* if the representation of the optimal w^* has a nonzero coefficient in $y_i x_i$; use the KKT condition to show that, any support vector i must have unnormalized margin $y_i(w^\star)^\top x_i \leq 1$.

(b) For each of the functions K below, state if it is a valid kernel function or not. If it is a kernel, write down its feature map; if not, prove that it is not one.

(1) $x = (x_1, x_2)$ and $z = (z_1, z_2)$ are real vectors; let $K(x, z) = x_1 \cdot z_2$.

(2) $x = (x_1, \ldots, x_d)$ and $z = (z_1, \ldots, z_d)$ are vectors whose entries are integers between 0 and 100; let $K(x, z) = \sum_{i=1}^{d} \min(x_i, z_d)$.

(3) $x = (x_1, \ldots, x_d)$ and $z = (z_1, \ldots, z_d)$ are real vectors; let $K(x, z) = (1 + x_1 z_1) \cdot \ldots \cdot (1 + x_d z_d)$.

Problem 6 (1pt). What do you plan to do for your final project?