HOMEWORK 5

>>NAME HERE<<
>>ID HERE<<

Instructions: Although this is a programming homework, you only need to hand in a pdf answer file. There is no need to submit the latex source or any code. You can choose any programming language, as long as you implement the algorithm from scratch.
Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. Please check Piazza for updates about the homework.

Linear Regression (100 pts total, 10 each)

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 $x = 2001, y = 21$.

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

2. Split the datasets: $x \leq 1970$ 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.

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

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

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

4. Using the MLE above, give the (1) mean squared error and (2) R^2 values on the Mendota test set. (You will need to use the Monona test data as observed features.)

5. "Reset" to Q3, but this time use gradient descent to learn the β's. Recall our objective function is the mean squared error on the training set:

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

Derive the gradient.

6. 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. Tell us if further iterations make your gradient descent converge, and if yes when; compare the β's to the closed-form solution. Try other 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.

7. As preprocessing, normalize your year and Monona features (but not $y_{Mendota}$). Then repeat Q6.

8. "Reset" to Q3 (no normalization, use closed-form solution), but train a linear regression model without using Monona:
$$\hat{y}_{Mendota} = \gamma_0 + \gamma_1 x.$$
   (a) Interpret the sign of γ_1.
   (b) Some analysts claim that because β_1 the closed-form solution in Q3 is positive, fixing all other factors, as the years go by the number of Mendota ice days will increase, namely the model in Q3 indicates a cooling trend. Discuss this viewpoint, relate it to question 8(a).

9. Of course, Weka has linear regression. Reset to Q3. Save the training data in .arff format for Weka. Use classifiers / functions / LinearRegression. Choose "Use training set." Bring up Linear Regression options, set "ridge" to 0 so it does not regularize. Run it and tell us the model: it is in the output in the form of "β_1 * year + β_2 * Monona + β_0."

10. Ridge regression.
   (a) Then set ridge to 1 and tell us the resulting Weka model.
   (b) Meanwhile, derive the closed-form solution in matrix form for the ridge regression problem:
$$\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} (x_i^\top \beta - y_i)^2 \right) + \lambda \|\beta\|_A^2$$
   where
$$\|\beta\|_A^2 := \beta^\top A \beta$$
   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, do not blindly copy lecture notes.
   (c) Let $\lambda = 1$ and tell us the value of β from your ridge regression model.

Extra Credit: Multinomial Naïve Bayes [10 pts]

Consider the Multinomial Naïve Bayes model. For each point $(\mathbf{x}, y)$, $y \in \{0, 1\}$, $\mathbf{x} = (x_1, x_2, \ldots, x_M)$ where each x_j is an integer from $\{1, 2, \ldots, K\}$ for $1 \le j \le M$. Here K and M are two fixed integer.
Suppose we have N data points $\{(\mathbf{x}^{(i)}, y^{(i)}) : 1 \le i \le N\}$, generated as follows.

for $i \in \{1, \ldots, N\}$:
  $y^{(i)} \sim$ Bernoulli(ϕ)
  for $j \in \{1, \ldots, M\}$:
    $x_j^{(i)} \sim$ Multinomial$(\theta_{y^{(i)}}, 1)$

Here $\phi \in \mathbb{R}$ and $\theta_k \in \mathbb{R}^K (k \in \{0, 1\}$ are parameters. Note that $\sum_l \theta_{k,l} = 1$ since they are the parameters of a multinomial distribution.
Derive the formula for estimating the parameters ϕ and θ_k, as we have done in the lecture for the Bernoulli Naïve Bayes model. Show the steps.

Extra Credit: Logistic Regression [10 pts]

(1) Suppose for each class $i \in \{1, \ldots, K\}$, the class-conditional density $p(\mathbf{x}|y = i)$ is normal with mean $\mu_i \in \mathbb{R}^d$ and the same covariance $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$:
$$p(\mathbf{x}|y = i) = N(\mathbf{x}|\mu_i, \mathbf{\Sigma}).$$
Compute $p(y = i|\mathbf{x})$. Can it be represented as a softmax over a linear transformation of $\mathbf{x}$? Show the calculation steps.

(2) Suppose $p(\mathbf{x}|y=i)$ has different covariances $\mathbf{\Sigma}_i \in \mathbb{R}^{d\times d}$:

$$p(\mathbf{x}|y=i) = N(\mathbf{x}|\mu_i, \mathbf{\Sigma}_i).$$

Again, compute $p(y=i|\mathbf{x})$. Can it be represented as a softmax over a linear transformation of $\mathbf{x}$? Show the calculation steps.