Homework 3

1. **Linear regression models with one-hot embeddings.** Suppose $u$ is a categorical that can take $k$ values, i.e., $\mathcal{U} = \{1, \ldots, k\}$. The one-hot embedding of $u$ into $\mathbb{R}^k$ is defined as $\phi(u) = e_u$, where $e_j$ is the $j$th unit vector. We will add a first feature which is constant, i.e., $x_1 = 1$, so the embedding we use is $x = \phi(u) = (1, e_u) \in \mathbb{R}^d$ with $d = k + 1$.

We have a data set with $n$ observations, $x^1, \ldots, x^n, y^1, \ldots, y^n$.

(a) Show that the data matrix $X \in \mathbb{R}^{n \times d}$ (with rows $(x^1)^T, \ldots, (x^n)^T$) always has linearly dependent columns. This means that we cannot use (basic) least squares to fit a regression model, when we use one-hot embedding of a categorical.

*Hint.* Consider the sum of last $k$ columns in $X$.

(b) Now suppose that we add quadratic regularization on $\theta_{2:k+1}$ to our fitting method (with $\lambda > 0$). We do not regularize the model coefficient associated with the constant feature $x_1 = 1$. Show that the associated least squares problem has linearly independent columns.

(c) Show that the sum of the last $k$ coefficients $\theta_i$ (i.e., those associated with $u$) is zero, i.e., $\sum_{i=2}^{k+1} \theta_i = 0$.

*Hint.* Consider the least squares problem

$$\text{minimize } \|Az - b\|,$$

where $z$ is the variable, and $A$ and $b$ are problem data, where $A$ has linearly independent columns. The least squares solution is $\hat{z} = (A^TA)^{-1}A^Tb$, and the optimal residual is $\hat{r} = A\hat{z} - b$. The *orthogonality principle* states that for any $z \in \mathbb{R}^n$, we have

$$(Az) \perp \hat{r}.$$  

(d) Verify the properties in parts (a), (b), and (c) numerically, with a small problem with a constant feature and a single categorical $u$ that takes three values. You can use a data set with $n = 10$ samples. Generate random $u^i \in \{1, 2, 3\}$ and $y^i \in \mathbb{R}$ and then form the data matrix $X$. Verify that the columns are dependent. Then find the model coefficients using quadratic regularization. Verify that the sum of the three model coefficients associated with $u$ (i.e., $\theta_2 + \theta_3 + \theta_4$) is zero.

*Julia hint.* `rand(1:3)` will generate a random number in $\{1, 2, 3\}$.

*Remark.* The simple ridge regression problem above, with $u$ consisting of one categorical, can be solved analytically. (We are not asking you to do this.) But the conclusions of parts (a)–(c) hold when the raw input $u$ contains other features in addition to the categorical.

2. **Fitting the leaf values in a tree model.** Consider a tree predictor, characterized by the tree, and at each non-leaf vertex, the feature to split on and the threshold, and at each
leaf, a value $\hat{y}$. There are many ways to fit a tree predictor to a data set, but we won’t explore these methods in this course. Here we consider a simpler problem, which has a simple solution.

Suppose the tree is fixed, that is, for each non-leaf vertex, we fix the feature to split on, and we fix the threshold. To fully specify the tree model, we need to give the value of $\hat{y}$ for each leaf vertex. We denote the leaf values as $\hat{y}_j$, for $j = 1, \ldots, p$, where $p$ is the number of leaves in the tree. We collect these leaf values into the parameter vector $\theta = (\hat{y}_1, \ldots, \hat{y}_p) \in \mathbb{R}^p$, and focus on how to choose $\theta$ using ERM with quadratic and absolute loss function, with data set $x^1, \ldots, x^n, y^1, \ldots, y^n$.

Define the index sets, for $j = 1, \ldots, p$,

$$I_j = \{i \mid x^i \text{ ends up in leaf } j\},$$

i.e., the indexes of the data points for which the tree predictor evaluates $x^i$ to leaf $j$ (and so, makes prediction $\hat{y} = \hat{y}_j$). We will assume that these are all non-empty; that is, for every leaf, there is at least one data point that ends up at that leaf.

(a) Explain how to choose $\theta$ using ERM with quadratic loss. You can give your answer in English, and without justification.

(b) Explain how to choose $\theta$ using ERM with absolute loss. You can give your answer in English, and without justification.

3. Temperature forecasting. We consider the problem of forecasting tomorrow’s temperature, from today’s and previous observed temperatures. (We will work with the minimum daily temperature, but that doesn’t matter to us.) This dataset gives $T_t$, $t = 1, \ldots, 3650$, the temperature on day $t$ over a 10 year period, 1981–1990. In temp.json, you will find a 2920-vector $T_{\text{train}}$ consisting of the data between 1981 and 1988, and a 730-vector $T_{\text{test}}$ consisting of the data between 1988 and 1990.

We will forecast the daily temperature of day $t + 1$, $T_{t+1}$, given the past $m$ days of temperatures $T_{t-m+1}, \ldots, T_t$. We refer to $m$ as the memory of the model. We will use a linear regression model of the form

$$\hat{T}_{t+1} = \theta_1 + \theta_2 T_t + \theta_3 T_{t-1} + \cdots + \theta_{m+1} T_{t-m+1}.$$ 

You will choose the parameters $\theta \in \mathbb{R}^{m+1}$ using ridge regression, regularizing only $\theta_2, \ldots, \theta_{m+1}$. Plot the test errors for regularization parameter taking 100 values logarithmically spaced between $10^{-2}$ and $10^3$. Report the best value of $\lambda$, and the associated RMS test error. Report the parameters $\theta$ that achieve the smallest RMS test error. For each case, interpret what the coefficients mean. Do this for the values $m = 0$, $m = 1$, and $m = 20$. In addition, for the $m = 20$ case, plot the coefficients $\theta_2, \ldots, \theta_{21}$.

Which is your best overall predictor, and what RMS test error does it achieve?

Remarks and hints.
• For $m = 0$ you are predicting the temperature as a constant; for $m = 1$ you are predicting tomorrow’s temperature as an affine function of today’s temperature; and for $m = 20$ you are predicting tomorrow’s temperature as an affine function of the temperature over the last ten days. Be sure that the coefficients you find make sense.

• Train and test your models on the same size data sets. For training, this means going from $t = 20$ to $t = 2919$. (For $m = 20$, the first prediction you can make is for $\hat{T}_{21}$.) For testing, this means evaluating the RMS error from $t = 20$ to $t = 729$.

*Julia hints.* To produce $n$ values uniformly spaced on a log scale between $10^a$ and $10^b$, use `10 .^ range(a, stop=b, length=n)`. To find the index of the minimum value in a vector $v$, use `argmin(v)`. 