Homework 3

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

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

   (a) Show that the data matrix $X \in \mathbf{R}^{n \times d}$ (with rows $(x^1)^T, \dots, (x^n)^T$) 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. If X has linearly dependent columns, there is a vector $z \neq 0$ such that $Xz = 0$ (z is in the nullspace of X). Try to construct this vector; you can use it in your answer.

   (b) Now suppose that we add ridge regularization (also called 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 θ_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 $\hat{z} = (A^T A)^{-1} A^T b$, and the optimal residual is $\hat{r} = A\hat{z} - b$. The *orthogonality principle* states that for any $z \in \mathbf{R}^n$, we have

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

   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 yet.) But the conclusions of parts (a)–(c) hold when the raw input u contains other features in addition to the categorical.

2. *Regularized least squares and features.* The following problem will use `U`, and `v` found in `prostate_cancer_data.json`. For the problem below, split the data using a 70-30 train-test split.

   (a) Explain how to formulate the problem of fitting regularized least squares given a matrix U and regularization parameter $\lambda > 0$, where the first feature is the constant feature $x_1 = 1$.

   (b) Standardize all of the features and then fit a model to the data, adding only a constant feature. Sweep your regularization parameter λ over the range $[10^{-5}, 10^5]$ and plot the corresponding training and test errors. Choose an appropriate value for λ, *i.e.*, the largest value that achieves approximately minimum test error. Give the model, and the corresponding test error.

   (c) If you look into the data matrix U, you'll notice that the last two columns actually take on only a few values. Embed both columns using a one-hot encoding, keeping the rest of the values the same. Now do the same sweep you did in part (b), and plot the corresponding training and testing errors. Compare the final test RMSE of the one-hot encoding vs. the original encoding, with appropriately chosen λ for each.

   Hint. You can make use of the Julia file `to_one_hot.jl`, which contains the function `to_one_hot(u)`. This function takes as input an n-vector u whose entries are one of k categories and embeds it using a one-hot embedding into $\mathbf{R}^{n\times k}$.

3. *Wildfire predictor.* Our task is to create a predictor which identifies the parts of the forest at risk from wildfires. We would like to predict the burned area given a set of features presented in the table below.

Features	**Description**	**Range**
position x	x-axis spatial coordinate within the park	1-9
position y	y-axis spatial coordinate within the park	1-9
month	month of the year	1-12
FFMC	Fine Fuel Moisture Code	18.7-96.2
temp	temperature in Celsius degrees	2.2-33.3
wind	wind speed in km/h	0.4-9.4
rain	outside rain in mm/m^2	0.0 to 6.4
Label	**Description**	**Range**
area	burned area of the forest in ha on a log-scale	0.00-6.99

The Fine Fuel Moisture Code (FFMC) is a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel.

In `wildfire_data.json`, you will find an $n \times d$ matrix `U` of data, with rows $(u^i)^T$, with $u^i \in \mathbf{R}^7$, and n-vector `v` with the burned area data. The columns are listed in order, so `position x` is first, followed by `position y`, followed by `month`, etc.

Randomly partition the data into a training set consisting of 80% of the data and a validation set consisting of the remaining 20% of the data. We will work with $y = \psi(v) = v$.

(a) Standardize the training set. Report the means and standard devations of each feature column before and after the standardization. For the validation set, standardize using the mean and standard deviation from the corresponding feature columns in the training set. Report the means and standard deviations of each feature column in the validation set before and after this transformation.

Hint: Import `Statistics` in your Julia file or notebook. Then you can compute the mean of the columns of a matrix `U_raw` as `mean(U_raw, dims=1)`. You can compute the standard deviation of the columns of a matrix `U_raw` as `std(U_raw, dims=1)`.

Hint: To standardize a matrix `U_raw`, compute `U_std = (U_raw .- U_mean) ./ U_std`

(b) For both the training and test sets, encode the month features as 12 one-hot embedded features. Standardize these features. Create a data matrix containing this embedded month feature along with the other six features and an constant (intercept) feature (a vector of ones). (There should be a total of 19 features.)

Print one row of your data matrix and verify it contains reasonable values.

(c) For 100 values of the regularization parameter λ uniformly spaced on a log scale between 10^{-1} and 10^{5}, fit a ridge regression model to the data. Plot the train and test RMS errors versus lambda using a log scale for the λ-axis.

Report the smallest RMS error you achieve and the corresponding value of λ that achieves it.

Hint: To produce n values uniformly spaced on a log scale between 10^a and 10^b, use `10 .^ range(a, stop=b, length=n)`.

Hint: Suppose you have a vector `losses` of RMSE values for different λ. `minimum(losses)` returns the minimum value. `argmin(losses)` returns the **index** of the minimum value..

(d) For both the training and test sets, transform the month features using sinusoidal embeddings: replace u_3^i by $(\sin(2\pi u_3^i/12),\ \cos(2\pi u_3^i/12))$. Using this embedding instead of the one-hot embedding, repeat the experiment in part (b). (Including the constant feature, the transformed data should have 9 features.)

(e) Provide a (short!) justification for the sinusoidal embedding. This can be one or two sentences.