50.021 Artificial Intelligence
Homework 1

Due: every Monday, 4PM before class starts

[Q1.] Write down the distribution of $p(x, y)$ from in class coding exercise. Note that $p(y = 0|x, c(x) = 1) = 0.8$ and $p(y = 0|x, c(x) = 2) = 0.7$. It is 60% more likely to draw samples from gaussian 1.

[Q2]. In the lecture notes, we solve the objective function:

$$\hat{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)^2 \tag{1}$$

by hand, and we have the analytical solution for optimum weights $\boldsymbol{w}*$ in linear regression,

$$\hat{\boldsymbol{w}} = (X^T \cdot X)^{-1} X^T \cdot Y$$

$$\hat{\boldsymbol{w}} \in \mathbb{R}^{d \times 1}, \boldsymbol{X} \in \mathbb{R}^{N \times d}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^{N \times 1},$$

Now suppose instead of using $\boldsymbol{x}_i$ directly, we want to use some mapping function $\phi(\boldsymbol{x}_i) = (\phi_1(\boldsymbol{x}_i), \ldots, \phi_C(\boldsymbol{x}_i))$ on each data sample x_i, and suppose that we use a slightly different squared error loss function than the lecture notes,

$$L(y, f(\boldsymbol{x})) = \frac{1}{2N} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2$$

$$f(\boldsymbol{x}) = \phi(\boldsymbol{x}) \cdot \boldsymbol{w}$$

1. Show that the solution for optimum weight $\hat{\boldsymbol{w}}$ still takes the similar form,

   $$\hat{\boldsymbol{w}} = (\Phi^T \cdot \Phi)^{-1} \Phi^T \cdot Y,$$

   where,

   $$\Phi = \begin{bmatrix} \phi_1(\boldsymbol{x}_1) & \ldots & \phi_C(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_N) & \ldots & \phi_C(\boldsymbol{x}_N) \end{bmatrix} \in \mathbb{R}^{N \times C},$$

   and N is the number of samples.

2. Show that if we define a function,

   $$\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_r) = \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_r),$$

   then $f(\boldsymbol{z})$ can be written only in terms of the function $\mathcal{K}(x_i, x_j), i, j = 1, \ldots, N$, the function $\mathcal{K}(\boldsymbol{z}, x_i)$ and Y without the need to specify ϕ explicitly. Write down the solution for $\boldsymbol{v}$ and $f(\boldsymbol{z})$ using that optimal $\boldsymbol{v}$.
   Hint: Use the assumption $\boldsymbol{w} = \Phi^T \boldsymbol{v}$, then optimize for the new parameter vector $\boldsymbol{v}$ to obtain a solution for $\boldsymbol{v}$ which does depend only on Y and K.

[Q3] Programming question. Specify instructions and the environment needed to run your code.

I. Linear Features:
Consider the two simple data sets *dataLinReg1D.txt* and *dataLinReg2D.txt*. The data files can be plotted using matplotlib. Each line contains a data entry (x, y) with $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. The last entry in each line refers to y. Compute and report the optimal parameters w for a linear Ridge regression model (just linear features) for both data sets.
Tips:

a) Write a routine that loads a data file and returns a matrix X containing all x_i as rows, and a vector y containing all y_i.

b) Write a routine that takes the raw X as input and returns a new X with a '1' pre-pended to each row. This routine simply computes the linear features including the constant '1'. This routine can later be replaced by others to work with non-linear features.

c) Write a routine that returns the optimal w from X and y - analytically, not by gradient descent.

d) Generate some test data points (along a grid) and collect them in a matrix Z. Apply routine b) to compute features. Compute the predictions $y = Zw$ (simple matrix multiplication) on the test data and plot it.

II. Cross-validation:
Implement 5-fold cross-validation to evaluate the generalization performance of the linear and rbf-basis function regression method for *dataLinReg2D.txt*.

Repeat the whole experiment starting from randomized cross-validation 10 times. Every time you choose an optimal lambda based on your cross-validated error $\hat{e}$. What is the distribution of the optimal λ now for this low-noise setting??

Report 1) the distribution of the optimal λ (e.g. by a histogram). Report 2) the mean squared error $\hat{e}$ from cross-validation, and 3) the standard deviation of $\hat{e}$ as a function of different Ridge regularization parameters, λ - start at $\lambda = 1e-4, \ldots, 10$ (ideally, generate a nice bar plot of the generalization error, including deviation, for various λ), averaged over the cross-validation folds and the ten time repetition.

Now you add to every label gaussian noise with standard deviation of 10.

$$y_i = y_i + \epsilon, \epsilon \sim N(0, 100)$$

Repeat the whole experiment again 10 times. Every time you choose an optimal lambda based on your cross-validated error $\hat{e}$. What is the distribution of the optimal λ now for this noisier setting?? Report 4) the distribution of the optimal λ (e.g. by a histogram).