CS559 Homework #1
Due: 2/11/2020 11:59 PM

Homework #1 is an individual work: each student must submit their own work. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

Probability Theory

1. By using a change of variables, verify that the univariate Gaussian distribution given by

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

satisfies $E(x) = \mu$. Next, by differentiating both sides of normalization condition

$$\int_{-\infty}^{-\infty} N(x|\mu, \sigma^2)dx = 1$$

with respect to σ^2, verify that the Gaussian satisfies $E(x^2) = \mu^2 + \sigma^2$.

2. Use $E(\boldsymbol{x}) = \boldsymbol{\mu}$ to prove $E(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$. Now, using the results two definitions, show that

$$E[x_n x_m] = \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{T}} + I_{nm}\boldsymbol{\Sigma}$$

where x_n denotes a data point same from a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and I_{nm} denotes the (n,m) element of the identity matrix. Hence prove the result (2.124)

Linear Regression

3. Consider a linear model of the form:

$$f(\mathrm{x}, \mathrm{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum of squares/loss function of the form:

$$L_D(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}(f(x_n, \mathrm{w}) - y_n)^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $E[\epsilon_i] = 0$ and $E[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, show that minimizing L_D averaged over the noise distribution is equivalent to minimizing the sum of square error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameters w_0 is omitted from the regularizer.

4. UCI Machine Learning: Bike Sharing Data Set
Build at least four regression models (e.g., linear, polynomial, non-linear) to predict the count of total rental bikes including both casual and registered. Explore data to reduce the number of features. Use K-fold cross validation and report the mean squared error (MSE) on the testing data. You need to write down every step in your experiment.

5. UCI Machine Learning: Iris Data set
a. Implement Linear Discriminant Analysis for each pair of the classes and report your results. Note that there are three class labels in the data set. Write down each step of your solution.

b. Perform the kNN classification for each k value from 1 to 50 to predict the species. For each k value, compute the percentage of misclassified values on the testing set. Print out your results as a table showing the values of k and the misclassification percentages. Then plot the misclassification rates on the testing set versus the k values.

Video Presentation. (Has a separate due date)

The assigned students below find a research paper related to assigned topics and make a short, between 10 and 15 minutes, video presentation. In the presentation, have the summary of paper, research question, method, and conclusion. Upload the file in the discussions panel before next week Tuesday 2/4th 11:59 PM. Rest students in class make questions or comments about the paper for each presentation by Tuesday 2/11th 11:59 PM. Then the presenter must make replies before Tuesday 2/18th.

Kelsey Douma - Exploratory Data Analysis / Missing Data Handling
David Etler - Linear Regression