2EL1730: MACHINE LEARNING
CENTRALESUPÉLEC

Assignment 1

Instructors: Fragkiskos Malliaros and Maria Vakalopoulou
TAs: Enzo Battistella, Yoann Pradat, Jun Zhu

Due: **December 23, 2020 at 23:00**

How to submit: Please complete the first assignment **individually**. *Typeset* all your answers (**PDF** file only). Submissions should be made on **gradescope** (Assignment 1; Entry Code: D5JPBB). Use your full name while registering in Gradescope (**FirstName LastName**). Make sure that the answer to each question is on a **separate page** (questions 1-8).

I. The Learning Problem, Model Selection and Evaluation

Question 1 [5 points]

Choose the correct answer *with justification*. *[Keep your answer short]*
Suppose you have picked the parameter(s) θ for a model using 10-fold cross validation (CV). The best way to pick a final model to use and estimate its error is to:

1. Pick any of the 10 models you built for your model; use its error estimate on the held-out data.
2. Pick any of the 10 models you built for your model; use the average CV error for the 10 models as its error estimate.
3. Average all of the 10 models you got; use the average CV error as its error estimate.
4. Average all of the 10 models you got; use the error the combined model gives on the full training set.
5. Train a new model on the full data set, using the θ you found; use the average CV error as its error estimate.

Question 2 [15 points]

Let's consider that we are trying to estimate a function $g(z) = 2z^2$, for $z \in \mathbb{R}$, using a linear regression model $h(z) = \theta^\top z$, based on a set of data points $x \in \mathbb{R}$ drawn from a uniform distribution $\mathcal{U}(-1, 1)$. Each point is also associated with a label $y = g(x)$ (the labels are noise-free). Let's assume that we train this linear model $h(\cdot)$ with *just a single data point* $x \neq 0$.

(a) [2 p] Explain briefly (1-2 lines) why do we expect that this linear model $h(z)$ will have large bias.

(b) [5 p] What is the bias of $h(z)$? This should be expressed as a function of a test data point $z \in \mathbb{R}$, without including any component related to x. (*Hint: first compute the regression coefficient* θ).

(c) [5 p] Similarly, compute the variance of the model as a function of a test data point $z \in \mathbb{R}$. The term computed should be a function of z (without including an x).

(d) [3 p] Let's assume that $Er(h, z)$ is the mean squared error for a test point $z \in \mathbb{R}$. Provide the relationship between between $Er(h, z)$, the bias and the variance at z. Then, compute precisely the values of each of the above three terms for $z = 1$.

II. Regression

Question 3 [5 points]

Choose the correct answer *with justification*. *[Keep your answer short]*
Suppose we have a regularized linear regression model: $\text{argmin}_w \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_p^p$. What is the effect of increasing p on bias and variance ($p \geq 1$), if the weights are all larger than 1?

1. Increases bias, increases variance.
2. Increases bias, decreases variance.
3. Decreases bias, increases variance.
4. Decreases bias, decreases variance.
5. Not enough information to tell.

Question 4 [20 points]

Let $\{y_i, X_i\}_{i=1}^m$ denotes a set of m observations, where each X_i is an n-dimensional vector. In *Ridge Regression*, a regularization term is added in the linear regression model in order to penalize the model complexity, leading to the following optimization problem (special case of the formula given in Question 7):

$$\arg\min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_2^2,$$

where $\lambda > 0$ is a regularization parameter.

(a) [12 p] Find the closed form solution of the ridge regression problem.

(b) [5 p] Calculate the Hessian matrix $\nabla_\theta^2 J(\theta)$, where $J(\theta)$ corresponds to the objective function.

(c) [3 p] Explain briefly why the ridge regression estimator is more robust to overfitting compared to the least-squares regression.

Question 5 [15 points]

Let $\tanh(\alpha) = \dfrac{e^\alpha - e^{-\alpha}}{e^\alpha + e^{-\alpha}}$ be the hyperbolic tangent function.

(a) [5 p] Show that the $\tanh(\cdot)$ function and the *logistic sigmoid function* $\sigma(\cdot)$ are related by:

$$\tanh(\alpha) = 2\sigma(2\alpha) - 1.$$

(b) [10 p] Show that a general (M-th order polynomial) linear combination of logistic sigmoid functions of the form

$$y(x, \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^{M} \theta_j \, \sigma\Big(\frac{x - \mu_j}{s}\Big)$$

is equivalent to a linear combination of $\tanh(\cdot)$ functions of the form

$$y(x, \mathbf{u}) = u_0 + \sum_{j=1}^{M} u_j \ \tanh\Big(\frac{x - \mu_j}{2s}\Big)$$

and find expressions to relate the new parameters $\{u_0, \ldots, u_M\}$ to the original parameters $\{\theta_0, \ldots, \theta_M\}$.

Question 6 [20 points]

In this exercise you will need to use the *Amazon Gift Card* dataset (`https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Gift_Card_v1_00.tsv.gz`). The above is a tab-separated values dataset, which includes reviews from Amazon products. You can import the data using a .csv reader of Python. Using the dataset, you will need to answer the following questions. You can use the `scikit-learn`[1] library for your models. **Include only the basic parts of your code in the report – Python scripts will not be submitted**.

(a) [2 p] What is the distribution of ratings in the dataset (e.g., number of 1-star, 2-star, 3-star (etc.) reviews)? Your answer can either be a table or a plot showing the distribution.

(b) [6 p] Now, we will train a simple *linear regression* model to predict the star rating of each review using just two features of the dataset:

$$\text{star\_rating} \simeq \theta_0 + \theta_1 \times \text{verified\_purchase} + \theta_2 \times \text{length of review},$$

where the 'length of review' is the number of words in the review (excluding ',' and '.'). Report the values of θ_0, θ_1, and θ_2 and briefly provide an interpretation of these values (i.e., what do they represent). Explain these in terms of the features and labels, e.g., if the coefficient of 'length of review' is negative, what would that say about positive versus negative reviews?

(c) [6 p] Split the dataset into two fractions: the first 90% of the dataset will be used for training, while the remaining 10% for testing (do not shuffle the dataset; use the order of the instances as they appear in the file). Now, train the same model as in question (b) on the training set, and report the model's mean-square error on both the training and test sets.

(d) [6 p] Consider a similar prediction problem as above, where you can use all four numeric features of the dataset ('verified purchase', 'length of review', 'helpful votes', 'total votes'). Try to obtain a more accurate model using *polynomial features*, as we have examined in the class[2]. Give the expression of the feature vector you have designed and report the training and test accuracies.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html`

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html`

V. Generative Models

Question 7 [10 points]

Consider the following probability distribution $P_\theta(x) = 2\theta x e^{-\theta x^2}$, where θ is a parameter and x is a positive real number. Suppose you get m *i.i.d.* samples x_i drawn from this distribution. Show how one can compute the *maximum likelihood estimator* for θ based on these samples.

VI. Nearest Neighbors

Question 8 [10 points]

Choose one or more correct answers *with justification [Keep your answer short]*

(a) [5 p] What tends to be true about increasing the *k* in *k*-nearest neighbors?

1. The decision boundary tends to get smoother.
2. The bias tends to increase.
3. The variance tends to increase.

(b) [5 p] What is the training error (fraction labeled wrong) in the below picture using 1-nearest neighbor and the $L1$ distance norm between points? *If there is more than one equally close nearest neighbor, use the majority label of all the equally close nearest neighbors. Please make sure you really are computing training error, not testing error.*

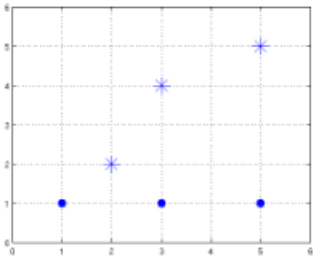

Figure 1: This picture, which shows a set of 6 points in a two-dimension space, where each point is either labeled "." and "*".

1. 0
2. $\frac{1}{6}$
3. $\frac{2}{6}$
4. $\frac{3}{6}$
5. $\frac{4}{6}$