# CS 529 - Assignment 1, Theory Solutions

## 1 Problem 1: Background

### 1.1 Part 1

- **Generalization**. Generalization refers to the ability of a machine learning model to achieve good performance when being applied to new, previously unseen data.
- **Overfitting and underfitting**. *Overfitting* refers to the situation where a machine learning model models the training data very well but perform badly on previously unseen testing data. Overfitting typically occurs when a model is very complex (i.e., has high variance) but the training data is limited. *Underfitting* refers to the situation where a model cannot model the training data well and also fails to generalize to new unseen data. Underfitting typically occurs when the model is too simple and so it cannot capture the nuances of the dataset of interest.
- **Regularization**. Regularization refers to a broad range of techniques to make learning algorithms favor more simpler models (i.e., it tries to combat the overfitting problem). For example, L1 and L2 regularization methods are commonly used to penalize models that have large weights.
- **No free lunch theorem**. The theorem states that there is no one single machine learning model that works best for every problem. For example, in a supervised learning problem where there are a lot of labeled examples, complex deep learning models may perform better than simple logistic regression model. On the other hand, when the data is limited, using deep learning models may result in overfitting and it may be better to use simpler models.
- **Occam's razor**. Occam's razor is a principle stating that given all other things being equal, simpler models should be favored over a more complex model.
- **Independent and identically distributed data points**. Independent and identically distributed data points are data points that (1) come from

the same distribution (2) and were sampled independently. In many supervised learning settings, it is typically assumed that the training / test examples are independent and identically distributed.

- **Cross-validation**. Cross-validation is a method for estimating a model performance on new dataset that it has not been trained on. There are several popular variants of cross-validation such as leave-one-out cross-validation or k-fold cross-validation. Cross-validation is also frequently used for hyper-parameter tuning.
- **Degrees of freedom**. Roughly speaking, the degrees of freedom refers to the number of values involved in some calculation that have the freedom to vary. In the context of supervised learning, typically, the more variables a model uses to predict a target, the more degrees of freedom the model has.

## 1.2 Part 2

Before tackling the main question, let's consider a related problem.

**MLE of a coin flip**. Suppose you have a coin whose probability of landing head is $\theta$. You toss the coin $N$ times in total and you observe $N_H$ heads and $N_T$ tails ($N = N_H + N_T$). Assume that the coin tosses are i.i.d. random variables. What is the Maximum Likelihood Estimate (MLE) of $\theta$?

Let $X_i$ denotes the outcome of the i-th toss. We set $X_i$ to be 1 if the outcome is head (and 0 otherwise). Let's start by calculating the likelihood:

$$\mathcal{L} = \prod_{i=1}^{N} \theta^{X_i}(1-\theta)^{(1-X_i)} \qquad \text{(The coin tosses are i.i.d)}$$

$$\mathcal{L} = \theta^{N_H}(1-\theta)^{N_T}$$

$$\log(\mathcal{L}) = N_H \log(\theta) + N_T \log(1-\theta) \qquad \text{(Taking log on both sides)}$$

$$\frac{\partial \log(\mathcal{L})}{\partial \theta} = \frac{N_H}{\theta} - \frac{N_T}{1-\theta} \qquad \text{(Taking derivative w.r.t } \theta\text{)}$$

In Maximum Likelihood Estimation (MLE), we basically need to find an estimate of the parameter $\theta$ that can maximize the (log) likelihood. Therefore, we can try setting $\frac{\partial \log(\mathcal{L})}{\partial \theta}$ to be 0. In other words:

$$\frac{\partial \log(\mathcal{L})}{\partial \theta} = 0 \Rightarrow \frac{N_H}{\theta} - \frac{N_T}{1-\theta} = 0$$

$$\Rightarrow N_H(1-\theta) = N_T\theta$$

$$\Rightarrow N_H = (N_H + N_T)\theta$$

$$\Rightarrow \hat{\theta} = \frac{N_H}{N_H + N_T} = \frac{N_H}{N}$$

In conclusion, in order to find the MLE for $\theta$, we basically just need to count the number of heads and then divide it by the total number of coin tosses.

Now let's get back to the main question.

- For the first coin, we have 12 heads out of 17 tosses. So the MLE for the probability of landing head for the first coin is $12/17 = 0.706$.
- For the second coin, we have 6 heads out of 17 tosses. So the MLE for the second coin is $6/17 = 0.353$.

So my guess for the first coin's next toss is H. And my guess for the second coin's next toss is T.

### 1.3 Part 3

Let's define the following matrix $A$:

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^M \\ 1 & x_2 & x_2^2 & \dots & x_2^M \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^M \end{bmatrix}$$

Basically, the i-th row of $A$ contains the bias term and input feature values for the i-th example in the dataset. In addition, let $\mathbf{w}$ be the column vector containing the weights of the model. In other word, $\mathbf{w} = \begin{bmatrix} w_0 & w_1 & \dots & w_M \end{bmatrix}^T$. Finally, let $\mathbf{t} = \begin{bmatrix} t_1 & t_2 & \dots & t_N \end{bmatrix}^T$ (i.e., $\mathbf{t}$ contains the true/target values for the examples).

With the newly defined vectors and matrix, we can rewrite the error function

$E(\mathbf{w})$ as follow:

$$E(\mathbf{w}) = \frac{1}{2}\left\|A\mathbf{w} - \mathbf{t}\right\|^2 = \frac{1}{2}(A\mathbf{w} - \mathbf{t})^T(A\mathbf{w} - \mathbf{t})$$

$$E(\mathbf{w}) = \frac{1}{2}\Bigg( \left\|A\mathbf{w}\right\|^2 + \left\|\mathbf{t}\right\|^2 - 2\mathbf{w}^T A^T t\Bigg)$$

Because we are interested in minimizing $E(\mathbf{w})$, let's take the derivative of $E(\mathbf{w})$ with respect to $\mathbf{w}$ (note that $\mathbf{w}$ is a vector so we need to use some basic matrix calculus knowledge). We have:

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{2}\left(\frac{\partial}{\partial \mathbf{w}}\left\|A\mathbf{w}\right\|^2 - 2\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^T A^T t\right)$$

$$= \frac{1}{2}(2A^T A\mathbf{w} - 2A^T\mathbf{t}) = A^T A\mathbf{w} - A^T\mathbf{t}$$

If we equate the derivative to 0, we can solve for $\mathbf{w}^\star$:

$$A^T A\mathbf{w}^\star - A^T\mathbf{t} = 0$$

$$\Rightarrow A^T A\mathbf{w}^\star = A^T\mathbf{t}$$

$$\Rightarrow \mathbf{w}^\star = (A^T A)^{-1}A^T\mathbf{t}$$

Thus we have found a closed-form solution $\mathbf{w}^\star$ for minimizing the error function. In fact, the above equation is commonly known as the normal equation [1].

Suppose in the future we are given an input $x$, then we can first construct the vector $\mathbf{x} = \begin{bmatrix}1 & x & x^2 & \dots & x^M\end{bmatrix}^T$. And then the prediction of the model will be:

$$y(x, \mathbf{w}^\star) = \mathbf{x}^T\mathbf{w}^\star = \mathbf{x}^T\Bigg((A^T A)^{-1}A^T\mathbf{t}\Bigg)$$

[1] Note that sometimes $A^T A$ may not be invertible. In that case, the system $A^T A\mathbf{w}^\star = A^T\mathbf{t}$ has more than one solution.

# 1 Question 2

1. **How many men and women (sex feature) are represented in this dataset?**
   Number of men in this dataset is: 21790
   Number of women in this dataset is: 10771

2. **What is the average age (age feature) of women?**
   The average age of women is: 36.86

3. **What is the percentage of German citizens (native-country feature)?**
   Percentage of German citizens: 0.42%

4. **What are the mean and standard deviation of age for those who earn more than 50K per year (salary feature) and those who earn less than 50K per year?**
   For people who earn less than 50K per year, the mean of age is 36.78 and the standard deviation of age is 14.02
   For people who earn more than 50K per year, the mean of age is 44.25 and the standard deviation of age is 10.52

5. **Is it true that people who earn more than 50K have at least high school education? (education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters or Doctorate feature)**
   False

6. **Display age statistics for each race (race feature) and each gender (sex feature).**
   Refer to Figure 1.

7. **What is the maximum number of hours a person works per week (hours-per-week feature)? How many people work such a number of hours, and what is the percentage of those who earn a lot (¿50K) among them?**
   Maximum number of hours a person works per week is: 99
   Number of people work such a number of hours is: 85
   The percentage of those who earn a lot (¿50K) among them is: 29.41%

8. **Count the average time of work (hours-per-week) for those who earn a little and a lot (salary) for each country (native-country). What will these be for Japan?**

For Japan
Average time of work for those who earn a little: 41.00
Average time of work for those who earn a lot: 47.96

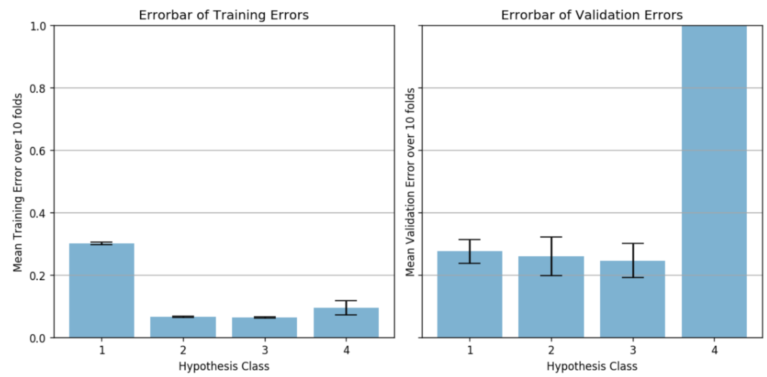
```
Age statistics for each race is shown below:
                     count       mean        std   min   25%   50%   75%  \
 race
 Amer-Indian-Eskimo    311.0  37.173633  12.447130  17.0  28.0  35.0  45.5
 Asian-Pac-Islander   1039.0  37.746872  12.825133  17.0  28.0  36.0  45.0
 Black                3124.0  37.767926  12.759290  17.0  28.0  36.0  46.0
 Other                 271.0  33.457565  11.538865  17.0  25.0  31.0  41.0
 White               27816.0  38.769881  13.782306  17.0  28.0  37.0  48.0

                     max
 race
 Amer-Indian-Eskimo  82.0
 Asian-Pac-Islander  90.0
 Black               90.0
 Other               77.0
 White               90.0


Age statistics for each gender is shown below:
           count       mean        std   min   25%   50%   75%   max
 sex
 Female  10771.0  36.858230  14.013697  17.0  25.0  35.0  46.0  90.0
 Male    21790.0  39.433547  13.370630  17.0  29.0  38.0  48.0  90.0


Age staistics for each race-gender combination is shown below:
                              count       mean        std   min   25%   50%  \
 race                sex
 Amer-Indian-Eskimo  Female    119.0  37.117647  13.114991  17.0  27.0  36.0
                     Male      192.0  37.208333  12.049563  17.0  28.0  35.0
 Asian-Pac-Islander  Female    346.0  35.089595  12.300845  17.0  25.0  33.0
                     Male      693.0  39.073593  12.883944  18.0  29.0  37.0
 Black               Female   1555.0  37.854019  12.637197  17.0  28.0  37.0
                     Male     1569.0  37.682600  12.882612  17.0  27.0  36.0
 Other               Female    109.0  31.678899  11.631599  17.0  23.0  29.0
                     Male      162.0  34.654321  11.355531  17.0  26.0  32.0
 White               Female   8642.0  36.811618  14.329093  17.0  25.0  35.0
                     Male    19174.0  39.652498  13.436029  17.0  29.0  38.0

                              75%   max
 race                sex
 Amer-Indian-Eskimo  Female  46.00  80.0
                     Male    45.00  82.0
 Asian-Pac-Islander  Female  43.75  75.0
                     Male    46.00  90.0
 Black               Female  46.00  90.0
                     Male    46.00  90.0
 Other               Female  39.00  74.0
                     Male    42.00  77.0
 White               Female  46.00  90.0
                     Male    49.00  90.0


The maximum age of men of Amer-Indian-Eskimo race is 82
```

Figure 1: Solution for Problem 2 Question 6

# 2 Question 3

**Plot the errorbar (i.e., the mean and std/sqrt(10)) of training and validation errors (you can use errorbar function, (Hint: x axis = 1,. . . , 4 (hypothesis class), y axis = mean error over 10 folds)).**
Refer to Figure 2.

**Plot the training input and outputs and the minimum training error hypothesis outputs for each hypothesis class above (4 plots, 10 hypotheses on each plot).**




Note: the validation error for Hypothesis 4 is too large to be plotted in the graph. To better compare the values, we listed the values in the table bellow.

| | Hypothesis 1 | Hypothesis 2 | Hypothesis 3 | Hypothesis 4 |
|---|---|---|---|---|
| mean(train error) | 0.300017 | 0.064747 | 0.063916 | 9.504845e-02 |
| std(train error) | 0.004397 | 0.002082 | 0.002047 | 2.288744e-02 |
| mean(validation error) | 0.274808 | 0.258657 | 0.245050 | 3.536696e+19 |
| std(validation error) | 0.038416 | 0.062015 | 0.054837 | 2.618335e+19 |

Figure 2: Solution for Problem 3 Question 4(a)

Refer to Figure 3.

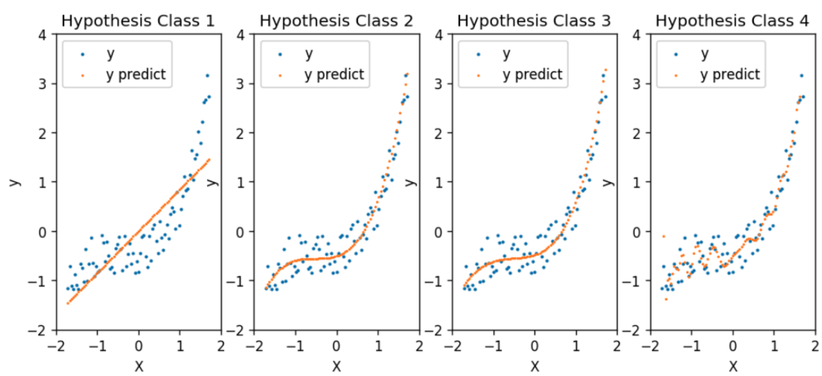



Figure 3: Solution for Problem 3 Question 4(b)

**Which hypothesis class would you choose among (a),. . . , (d) and why**
By Occam's razor principle, we can choose the relatively simple model from hypothesis 2/3.