HOMEWORK 1 - V1

CSC311 Spring 2020

- **Deadline:** Monday, February 3, 2020 at **16:59**.
- **Submission:** You need to submit two files through MarkUs. One is a PDF file including all your answers and plots. The other is a source file that reproduces your answers. You can produce the file however you like (e.g. LaTeX, Microsoft Word, etc) as long as it is readable. Points will be deducted if we have a hard time reading your solutions or understanding the structure of your code. If the code does not run, you may lose most/all of your points for that question.
- **Late Submission:** 10% of the marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.
- **Collaboration:** Weekly homework assignments must be done individually, and you cannot collaborate with others.

1. Nearest Neighbours and the Curse of Dimensionality – 25 pts. In this question, you will verify the claim from lecture that "most" points in a high-dimensional space are far away from each other, and also approximately the same distance.

(a) **[10 pts]** Consider two independent univariate random variables X and Y sampled uniformly from the unit interval $[0, 1]$. Determine the expectation and variance of the random variable $Z = |X - Y|^2$, i.e., the squared distance between X and Y.
Note: You can either compute the integrals yourself or use the properties of certain probability distributions. In the latter case, explicitly mention what properties you have used.

(b) **[10 pts]** Now suppose we draw two d-dimensional points X and Y from a d-dimensional unit cube with a uniform distribution, i.e., $X, Y \in [0, 1]^d$. Observe that each coordinate is sampled independently and uniformly from $[0, 1]$, that is, we can view this as drawing random variables $X_1, \ldots, X_d$ and $Y_1, \ldots, Y_d$ independently and uniformly from $[0, 1]$. The squared Euclidean distance $\|X - Y\|_2^2$ can be written as $R = Z_1 + \cdots + Z_d$, where $Z_i = |X_i - Y_i|^2$. Using the properties of expectation and variance, determine $\mathbb{E}\left[\|X - Y\|_2^2\right] = \mathbb{E}[R]$ and $\mathrm{Var}[\|X - Y\|_2^2] = \mathrm{Var}[R]$. You may give your answer in terms of the dimension d, and $\mathbb{E}[Z]$ and $\mathrm{Var}[Z]$ (the answers from part (a)).

(c) **[5 pts]** Based on your answer to part (b), compare the mean and standard deviation of $\|X - Y\|_2^2$ to the maximum possible squared Euclidean distance between two points within the d-dimensional unit cube (this would be the distance between opposite corners of the cube). Why does this support the claim that in high dimensions, "most points are far away, and approximately the same distance"?

2. Information Theory – 25 pts. The goal of this question is to help you become more familiar with the basic equalities and inequalities of information theory. They appear in many contexts in machine learning and elsewhere, so having some experience with them is quite helpful. We review some concepts from information theory, and ask you a few questions.

Recall the definition of the entropy of a discrete random variable X with probability mass function p:

$$H(X) = \sum_x p(x) \log_2 \left(\frac{1}{p(x)} \right).$$

Here the summation is over all possible values of $x \in \mathcal{X}$, which (for simplicity) we assume is finite. For example, $\mathcal{X}$ might be $\{1, 2, \ldots, N\}$.

(a) **[5pt]** Prove that the entropy $H(X)$ is non-negative.
(b) **[5pt]** If X and Y are independent random variables, show that $H(X, Y) = H(X) + H(Y)$
(c) **[5pt]** Prove the chair rule for entropy: $H(X, Y) = H(X) + H(Y|X)$.

An important concept in information theory is the relative entropy or the KL-divergence of two distributions p and q. It is defined as

$$\mathsf{KL}(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}.$$

The KL-divergence is one of the most commonly used measure of difference (or divergence) between two distributions, and it regularly appears in information theory, machine learning, and statistics. For this question, you may assume $p(x) > 0$ and $q(x) > 0$ for all x.

If two distributions are close to each other, their KL divergence is small. If they are exactly the same, their KL divergence is zero. KL divergence is not a true distance metric (since it isn't symmetric and doesn't satisfy the triangle inequality), but we often use it as a measure of dissimilarity between two probability distributions.

(d) **[5pt]** Prove that $\mathsf{KL}(p||q)$ is non-negative. *Hint: you may want to use Jensen's Inequality, which is described in the Appendix.*
(e) **[5pt]** The Information Gain or Mutual Information between X and Y is $I(Y; X) = H(Y) - H(Y|X)$. Show that

$$I(Y; X) = \mathsf{KL}(p(x, y)||p(x)p(y)),$$

where $p(x)$ is the marginal distribution of X and $p(y)$ is the marginal distribution of Y.

3. Decision Trees and K-Nearest Neighbour – 50 pts. In this question, you will use the `scikit-learn`'s decision tree and KNN classifiers to classify real vs. fake news headlines. The aim of this question is for you to read the `scikit-learn` API and get comfortable with training/validation splits.

We will use a dataset of 1298 "fake news" headlines (which mostly include headlines of articles classified as biased, etc.) and 1968 "real" news headlines, where the "fake news" headlines are from https://www.kaggle.com/mrisdal/fake-news/data and "real news" headlines are from https://www.kaggle.com/therohk/million-headlines. The data were cleaned by removing words from titles not part of the headlines, removing special characters and restricting real news headlines after October 2016 using the word "trump". The cleaned data are available as `clean_real.txt` and `clean_fake.txt` on the course webpage. It is expected that you use these cleaned data sources for this assignment.

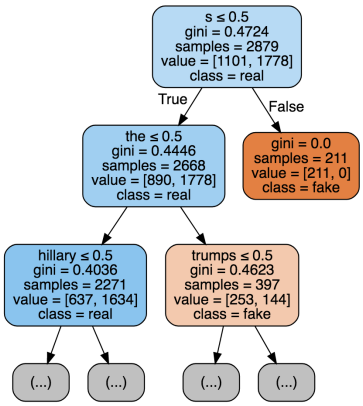
You will build a decision tree and KNN to classify real vs. fake news headlines. Instead of coding these methods yourself, you will do what we normally do in practice: use an existing implementation. You should use the `DecisionTreeClassifier` and `KNeighborsClassifier` included in `scikit-learn`. Note that figuring out how to use this implementation, its corresponding attributes and methods is a part of the assignment.

All code should be submitted in `hw1_code.py`.

(a) **[10 pts]** Write a function `load_data` which loads the data, preprocesses it using a vectorizer (http://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction.text, we suggest you use `CountVectorizer` as it is the simplest in nature), and splits the entire dataset randomly into 70% training, 15% validation, and 15% test examples. You may use `train_test_split` function of `scikit-learn` within this function.

(b) **[10 pts]** (Decision Tree) Write a function `select_tree_model` that trains the decision tree classifier using at least 5 different *sensible* values of `max_depth`, as well as two different split criteria (Information Gain and Gini coefficient), evaluates the performance of each one on the validation set, and prints the resulting accuracies of each model.
You should use `DecisionTreeClassifier`, but you should write the validation code yourself. Include the output of this function in your solution.

(c) **[10 pts]** (Decision Tree) Now let's stick with the hyperparameters which achieved the highest validation accuracy. Report its accuracy on the test dataset. Moreover, extract and visualize the first two layers of the tree. Your visualization may look something like what is shown below, but it does not have to be an image; it is perfectly fine to display text. It may also be hand-drawn. Include your visualization in your solution pdf.



(d) **[10 pts]** (Decision Tree) Write a function `compute_information_gain` which computes the information gain of a split on the training data. That is, compute $I(Y, x_i)$, where Y is the random variable signifying whether the headline is real or fake, and x_i is the keyword chosen for the split. Your split should be based on whether the keyword x_i exists (True) or does not exist (False). You should ignore the number of times that the keyword appears in the sentence.

Report the outputs of this function for the topmost split from the previous part, and for several other keywords.

(e) **[10 pts]** (KNN) Write a function `select_knn_model` that uses a KNN classifier to classify between real vs. fake news. Use a range of k values between 1 to 20 and compute both training and validation errors. You should generate a graph similar to the one on slide 43 of Lecture #1, which is Figure 2.4. of the Elements of Statistical Learning. You do not need to worry about the Bayes error or the Linear classifier in that figure. Report the generated graph in your report. Choose the model with the best validation accuracy and report its accuracy on the test data.

APPENDIX A: CONVEXITY AND JENSEN'S INEQUALITY

Here, we give some background on convexity which you may find useful for some of the questions in this assignment. You may assume anything given here.

Convexity is an important concept in mathematics with many uses in machine learning. We briefly define convex set and function and some of their properties here. Using these properties are useful in solving some of the questions in the rest of this homework. If you are interested to know more about convexity, refer to Boyd and Vandenberghe, Convex Optimization, 2004.

A set C is *convex* if the line segment between any two points in C lies within C, i.e., if for any $x_1, x_2 \in C$ and for any $0 \leq \lambda \leq 1$, we have

$$\lambda x_1 + (1 - \lambda)x_2 \in C.$$

For example, a cube or sphere in $\mathbb{R}^d$ are convex sets, but a cross (a shape like X) is not.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if its domain is a convex set and if for all x_1, x_2 in its domain, and for any $0 \leq \lambda \leq 1$, we have

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

This inequality means that the line segment between $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies above the graph of f. A convex function looks like $\smile$. We say that f is *concave* if $-f$ is convex. A concave function looks like $\frown$.

Some examples of convex and concave functions are (you do not need to use most of them in your homework, but knowing them is useful):

- Powers: x^p is convex on the set of positive real numbers when $p \geq 1$ or $p \leq 0$. It is concave for $0 \leq p \leq 1$.
- Exponential: e^{ax} is convex on $\mathbb{R}$, for any $a \in \mathbb{R}$.
- Logarithm: $\log(x)$ is concave on the set of positive real numbers.
- Norms: Every norm on $\mathbb{R}^d$ is convex.
- Max function: $f(x) = \max\{x_1, x_2, \ldots, x_d\}$ is convex on $\mathbb{R}^d$.
- Log-sum-exp: The function $f(x) = \log(e^{x_1} + \ldots + e^{x_d})$ is convex on $\mathbb{R}^d$.

An important property of convex and concave functions, which you may need to use in your homework, is *Jensen's inequality*. Jensen's inequality states that if $\phi(x)$ is a convex function of x, we have

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

In words, if we apply a convex function to the expectation of a random variable, it is less than or equal to the expected value of that convex function when its argument is the random variable. If the function is concave, the direction of the inequality is reversed.

Jensen's inequality has a physical interpretation: Consider a set $\mathcal{X} = \{x_1, \ldots, x_N\}$ of points on $\mathbb{R}$. Corresponding to each point, we have a probability $p(x_i)$. If we interpret the probability as mass, and we put an object with mass $p(x_i)$ at location $(x_i, \phi(x_i))$, then the centre of gravity of these objects, which is in $\mathbb{R}^2$, is located at the point $(\mathbb{E}[X], \mathbb{E}[\phi(X)])$. If ϕ is convex $\smile$, the centre of gravity lies above the curve $x \mapsto \phi(x)$, and vice versa for a concave function $\frown$.