CS37300 Homework	— Homework 4—	Last update: October 17, 2017
https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2017		Due: Oct 23, 2017 (6:00am)

Instructions and Policy: Each student should write up their own solutions independently. You need to indicate the names of the people you discussed a problem with; ideally you should discuss with no more than two other people.

- YOU MUST INCLUDE YOUR NAME IN THE HOMEWORK
- The answers MUST be in submitted in a single PDF file via Blackboard.
- Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary.
- Theoretical questions MUST include the intermediate steps to the final answer.

There are TWO (2) questions in this homework.

Q1 (50 pts): (Classification)

(a) **(30 pts)** (K-nearest neighbors [k-NN]) Consider a training dataset with consumer attributes (electricity consumption) E∈{High,Low}, (season) S∈{Winter, Summer, Spring, Fall}, (heating system) H∈{Oil,Electric}.

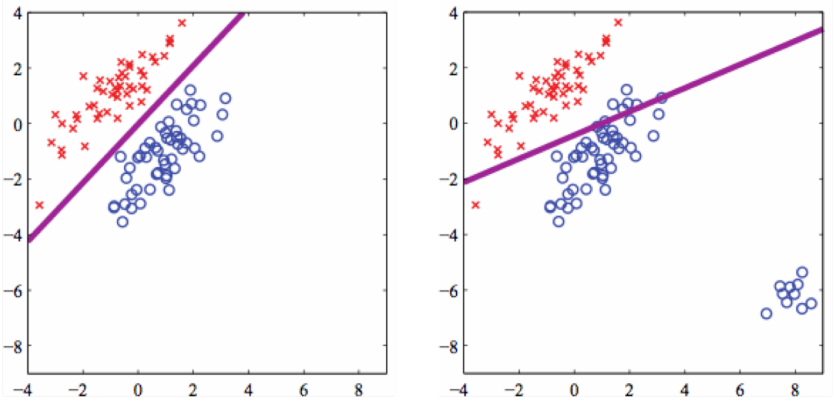
(i) **(10 pts, pseudocode)** Give the pseudocode of the k-NN algorithm to predict electricity consumption. Use the Hamming distance as your distance function. The Hamming distance counts the number of attribute values that are different in between two examples.

(ii) **(10 pts)** Using the 2-NN classifier, predict the electricity consumption of a new customer (in the test data) for season Winter and with heating system Electric using the training data in the table below. Describe which are the two nearest neighbors and what is the predicted consumption.

Electricity use	*Season*	*Heating system*
High	Winter	Oil
Low	Fall	Oil
High	Winter	Electric
High	Winter	Electric
Low	Summer	Oil
Low	Summer	Electric
High	Summer	Electric
Low	Spring	Electric

(iii) **(10 pts)** By increasing k, will the k-nearest neighbors be more likely to overfit the training data? Yes, no? Describe why using the table in item (ii) as an example.

(b) **(20 pts)** Consider the following binary classification problem, where the vertical and horizontal axis represent data attributes and the "+" or "O" sign represent distinct data labels.



(i) **(10 pts)** The above figures represent the decision boundaries of a linear classifier using the square loss. Describe why the figure on the right has a worse decision boundary than the figure on the left. Give a numerical example.

(ii) **(10 pts)** Describe two classifiers that do not suffer from the above issue.

Q2 (50 pts): (Feature construction) In this part of the homework we will learn how to construct features (classifier inputs) from categorical attributes. Consider a single neuron with a logistic activation function (Logistic regression) for the electricity consumption prediction problem described above.

A single logistic neuron has the following definition: let $\mathbf{x}_i \in \mathbb{R}^n$ be an n-dimensional vector representing the i-th customer. Let $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ be the parameters of the model. The probability a customer has high electricity cost is then defined as

$$P[y_i = \text{High}|\mathbf{x}_i] = \sigma(\mathbf{w}^T\mathbf{x}_i + b),$$

where

$$\sigma(a) = \frac{\exp(a)}{1+\exp(a)}.$$

Note that in the original problem all attributes are categorical, while in the single logistic neuron we need the attributes to be real-valued. In the following exercises we will learn how to transform the categorical attributes into real-value attributes.

(a) **(25 pts)** Consider two alternatives to transform the categorical attribute *Season* of Q1 into a real-valued attribute.
(Alternative a) Consider the first "categorical to real" mapping: Winter=0, Summer=1, Fall=2, Spring=4.
(Alternative b) Consider the second "categorical to real" mapping: Winter=-1, Spring=0, Fall=0, Summer=1.

In practice, we find that (Alternative a) has worse test data accuracy than (Alternative b). Can you explain why (Alternative b) is better than (Alternative a)?
Hint: Let $\mathbf{w}_S$ be the parameter in $\mathbf{w}$ that multiplies the season value. Use $\mathbf{w}_S$ to help explain the reason.
Hint2: Note that the fact (Alternative b) has negative values and (Alternative a) does not, does not mean much. The value of the intercept b can compensate these discrepancies.

(b) **(25 pts)** In item (a) you have seen that (Alternative b) was better than (Alternative a). Often, we don't know how the different categories are related (e.g., Summer is the opposite of Winter). Consider **(Alternative c)**, a "categorical to real vector" mapping known as *one-hot encoding*: Winter=(1,0,0,0), Summer=(0,1,0,0), Fall=(0,0,1,0), Spring=(0,0,0,1). Now, the old real value $\mathbf{w}_S$ becomes a 4-dimensional vector $\mathbf{w}_{S,\text{one-hot}}$. Explain why **(Alternative c)** is more flexible than **(Alternative b)**, i.e., show that there is a set of parameter $\mathbf{w}_{S,\text{one-hot}}$ that give the exact same behavior as **(Alternative b)**. Argue why **(Alternative c)** is better if we do not know how the categorical variables are related.