# CMPS142-Spring 2018
# Homework 1

Handed out: April 14, 2018
Due: April 26, 2018 at 1:30 PM

---

- You are allowed to solve this homework in groups of 2 or 3. Collaborating with any one not enrolled in the class, (except the course staff), or taking help from any online resources for the homework problems is strictly forbidden.
- One (and only one) member of the group has to submit the homework using his/her account on canvas. All group members will get points for that submission.
- How to submit your solutions: Your group's solution to each problem must be typed up separately (in at least an 11-point font) and submitted in the appropriate 'Problem' box on the Canvas website as a PDF file. **This means that if the homework has $N$ problems, you will submit $N$ separate pdf files on Canvas, one for each problem per group!** For example, submit the pdf that contains your group's solution to the first problem in the box titled 'Problem 1'. If a problem requires submitting additional files (like a zip file containing all your code or your processed data files) you will submit these along with the pdf for the problem.
- **Each pdf file** should clearly mention the names, email addresses and student ids of all group members. If you forget a group member's name, they will not get points for that problem.
- You are very strongly encouraged, but not required, to format your solutions in LATEX. You can use other softwares but handwritten solutions are not acceptable.
- Please try to keep the solution brief and clear.
- The homework is due at 1:30PM on the due date. There is a 10% penalty for each late day, upto 3 days. After that you will not get any points for this homework.
- The Computer Science Department of UCSC has a zero tolerance policy for any incident of academic dishonesty. If cheating occurs, consequences within the context of the course may range from getting zero on a particular homework, to failing the course. In addition, every case of academic dishonesty will be referred to the student's college Provost, who sets in motion an official disciplinary process. Cheating in any part of the course may lead to failing the course and suspension or dismissal from the university.

---

## Problem 1: Probability [10 points]

1. [5 points] A couple has two children and the older child is a boy. If the probabilities of having a boy or a girl are both 1/2, what is the probability that the couple has two boys?
2. [5 points] A couple has two children, of which at least one is a boy. If the probabilities of having a boy or a girl are both 1/2, what is the probability that the couple has two boys? *Hint: The answer to this question is not the same as the answer to the previous one.*

## Problem 2: Naive Bayes [10 points]

In this question, we'll discuss how to estimate the parameters using MLE for Naive Bayes.

Let $X = \langle X_1, X_2 \dots X_n \rangle$ be a vector representing a data instance. Each $X_i$ represents the value of the $i^{th}$ feature for $X$ and can take real values. All $X_i$'s are conditionally independent given the label. We model $P(X1 = x_{1j}|Y = y_k)$ using a Normal distribution as:

$$P(X_1 = x_{1j}|Y = y_k) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-(x_{1j} - \mu_{1k})^2}{2\sigma^2}\right)$$

Here $j \in \{1, 2, 3, ...M\}$ represents the $j^{th}$ training instance out of a total of $M$ instances. All instances are iid. Also, $x_{1j}$ refers to the value of the first attribute, $X_1$ of the $j^{th}$ instance. Note that here we are assuming that all classes, $k$, have the same variance, $\sigma$. In this question, we will estimate the parameter (mean of the Gaussian) for the first attribute. What is the maximum likelihood estimate for $\hat{\mu}_{1k}$? Show your derivation.

## Problem 3: Decision Trees [25 points]

Mary, the manager of a mattress store, collected a small dataset of her customer's attributes and the size of the mattress they purchased. The customers only bought two types of mattresses: King (K) or Queen (Q). Here is the data that Mary collected:

| Gender | Height | Preference |
|---|---|---|
| M | 5.2 | Q |
| M | 6.2 | Q |
| M | 6.8 | Q |
| M | 6.9 | K |
| M | 6.1 | K |
| F | 5.3 | K |
| F | 6.2 | Q |

Note that one of the attributes, Height, is a continuous variable. Lets assume that we will only allow binary splits for this attribute of the form Height$< h$ and Height$\geq h$, where $h$ lies in the dataset. However, there can be multiple such splits in one path from root to leaf.

1. [1 points] Lets say Mary wants to create a decision tree to predict the mattress type that a new customer will prefer. She wants to keep 'Height' at the root node. How many possible values of $h$ does she need to consider?
2. [3 points] What is the entropy of labels (mattress type) in the training dataset?
3. [5 points] What is the optimal root node for this dataset? Show your calculations.
4. [10 points] Draw the DT that would be learned by ID3 on this dataset? Also, label each non-leaf node with the gain attained by the corresponding split. How to submit: It is okay to draw the tree by hand and include a *clear* picture in your pdf.
5. [2 points] ID3 is a very popular algorithm for learning a decision tree. Does it learn an optimal tree? An optimal tree is one that has minimal depth and perfectly classifies the training data. Provide a short explanation for your answer.
6. [4 points] Change one attribute of one example in the given dataset, so that the learned tree will contain at least one more node. As an answer to this question, provide the new training dataset (highlighting the change), and the new learned tree.

## Problem 4: KL Divergence [5 points]

In this question, we explore the relationship between Information Gain, KL Divergence and Entropy. One way to understand KL divergence is to view it as a measure to estimate distance of a probability distribution, $p(x)$, from another probability distribution, $q(x)$. It is defined as:

$$KL(p||q) = -\sum p(x) log_2 \frac{q(x)}{p(x)}$$

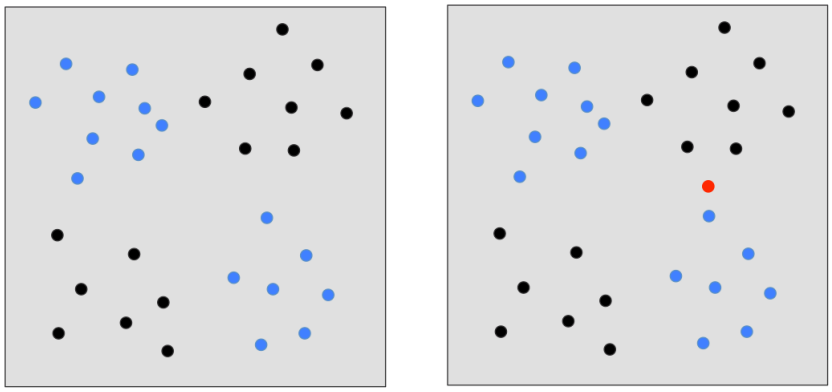
It is possible to define Information Gain (IG) as the KL-divergence from the product of the observed marginals of X and Y to their observed joint distribution.

$$IG(x,y) = KL(p(x,y)||p(x)p(y)) = -\sum_x \sum_y p(x,y) log_2 \frac{p(x)p(y)}{p(x,y)}$$

We, however, learned a different definition of IG in class: $IG(x,y) = H[x] - H[x|y] = H[y] - H[y|x]$. Show that this definition of IG is same as its definition in terms of KL-divergence.

## Problem 5: KNN [10 points]

In this problem, you will strengthen your understanding of the KNN algorithm. Each of the following figures shows a dataset. Each dataset contains examples from two classes, black and blue.



(a) Question about Training in KNN (b) Question about Testing in KNN

Figure 1: Figures for question about KNN

1. [4 points] Consider Figure 1(a) for this part of the question. Draw the decision boundaries of 1NN classifier for the datasets. Assume that the classifier works with Euclidean distance. You don't need to be super precise.
2. [6 points] Consider Figure 1(b) for this part of the question. What label would a KNN classifier predict for the red point if (a) K=1, and (b) K=3?

## Problem 6: Training DTs [40 points]

In this question, you will learn how to use a Machine Learning toolkit, Weka. Specifically, we will be learn Decision Trees for predicting survival using the Titanic dataset. It contains personal information about passengers and their tickets along with whether they survived. We have already split the data into train and test files: cmps142_hw1_train.csv and cmps142_hw1_test.csv. If interested, you could learn more about the corpus on its **Description** page. You need to learn using Weka (3.8.2) to train a Decision Tree classifier. The download page of Weka is **here**. A tutorial of how to use Weka to train a decision tree classifier is **here**. For this problem, we treat the 'survived' column of the corpus as label to be predicted.

1. [4 points] **Preprocessing:** The corpus has several attributes. By default, Weka assumes that most of them are numeric. However, some of the attributes would be meaningful as categorical/nominal attributes. In the first step, you will use Weka's inbuilt filter to convert these attributes from numeric to nominal for both the train and the test files. Specifically, you will use the filter named *NumericToNominal* to convert the following attributes to nominal: *Pclass, Parch, Has_Cabin, IsAlone, Survived*. After you have converted the above mentioned attributes to nominals for both the given files, save the resulting files as arff files named cmps142_hw1_train.arff and cmps142_hw1_test.arff respectively. Note that this can be done on the Weka GUI. You can read more about the arff format **here**.
   **What to submit:** You have to submit the two arff files with your report.
   **Questions to answer in report:** Using the visualization tool in Weka's GUI answer the following questions:

(a) How many unique values can the *Parch* attribute take in the train set? What are they?
(b) How many unique values can the *Parch* attribute take in the test set? What are they?
(c) [EXTRA CREDIT 2 points] Note that Weka automatically recognizes as *Sex*, *Embarked* and *Title* as Nominal attributes. Why does Weka think that an attribute like *Embarked* is nominal while an attribute like *Parch* isn't?

2. [8 points] **Building a tree:** For the rest of this homework, you will only work with the preprocessed arff files that you created in the previous step. Use C4.5 (J48) Decision Tree algorithm with Weka's default settings to learn a Decision Tree classifier on the training set. Note that the default settings are: C=0.25, M=2, and unpruned=False. For this homework assignment we are only concerned with these 3 parameters. The rest of the parameters/options should not be changed. We will be using 10-fold Cross-validation (CV) on the train set to evaluate the learned decision tree.
**Questions to answer in report:** Answer the following questions using Weka's GUI.
   (a) What is the 10-fold CV accuracy?
   (b) What is the confusion matrix for your 10-fold CV?
   (c) What is the (i) number of leaves and (ii) tree size as reported by Weka?

3. [10 points] **Pruning:** The tree built in the previous question was pruned. Now you will build an unpruned tree and analyze the difference in performance. To train an unpruned tree, use C4.5 (J48) Decision Tree algorithm with the following settings: M=2, and unpruned=True (setting unpruned=True renders C inactive). As before, report 10-fold CV performance on the train set.
**Questions to answer in report:**
   (a) What is the 10-fold CV accuracy?
   (b) What is the (i) number of leaves and (ii) tree size as reported by Weka?
   (c) How does the performance of the unpruned method compare with the performance using pruning? Also, give a reason for your observation.

4. [18 points] **Effect of pruning:** For this question, you will be training the tree using the train set and reporting performance on the training set itself as well as separately provided test set. Note that Weka allows you to do both. Weka's implementation of the algorithm has a parameter, C *confidenceFactor*, which controls the degree to which the tree is pruned. Smaller values of C result in more pruned trees. In this experiment, you will see the change in training and test performance with changing values of this parameter. Also, the value of M should be set to be 0.
**Questions to answer in report:**
   (a) Draw a plot of C versus training and test accuracies. The x-axis should report the following values of C {0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, and the y-axis should be training (and test accuracies). Report the trend that you observe. Why do you see this trend? Do not forget to include the plot in your submission.
   (b) Draw a plot of C versus size of tree. The x-axis should report the following values of C {0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, and the y-axis should be the corresponding tree size. Report the trend that you observe. Why do you see this trend? Do not forget to include the plot in your submission.

5. [EXTRA CREDIT 8 points] This question is similar to the previous one. In this question, you will study the effect of varying M on the performance on the train and the test sets. Plot the same graphs as above while trying the following values of M: {1,2,3,4,5,10,20}. Set unpruned=True. Report the trends you observe and an explanation for the trends. Do not forget to include the plots in your submission.