Simon Fraser University
Assignment 2 - CMPT 741 — Data Mining, Fall 2019

Date**: October 13th, 2019**

Name: **Anurag Bejju**
Student ID: **301369375**

---

1. Proof:

1.1. Given:

Good Class	Bad Class
$P(Creditworthiness = good) = 0.4$	$P(Creditworthiness = bad) = 0.6$
$P(Age = young \mid C = good) = 0.5$ $P(Age = medium \mid C = good) = 0.0$ $P(Age = old \mid C = good) = 0.5$	$P(Age = young \mid C = bad) = 0.5$ $P(Age = medium \mid C = bad) = 0.34$ $P(Age = old \mid C = bad) = 0.16$
$P(Salary = low \mid C = good) = 0.0$ $P(Salary = medium \mid C = good) = 0.75$ $P(Salary = high \mid C = good) = 0.25$	$P(Salary = low \mid C = bad) = 0.5$ $P(Salary = medium \mid C = bad) = 0.34$ $P(Salary = high \mid C = bad) = 0.16$
$P(City = Vancouver \mid C = good) = 0.25$ $P(City = Burnaby \mid C = good) = 0.0$ $P(City = Coquitlam \mid C = good) = 0.5$ $P(City = Richmond \mid C = good) = 0.25$	$P(City = Vancouver \mid C = bad) = 0.5$ $P(City = Burnaby \mid C = bad) = 0.34$ $P(City = Coquitlam \mid C = bad) = 0.0$ $P(City = Richmond \mid C = bad) = 0.16$

Required: Predict the class label (Creditworthiness) using Naïve Bayes classifier for an unseen client with Age = young, Salary = low and City = Vancouver

We need to calculate two conditional probabilities:

$$P(Creditworthiness = good \mid age = young,\ salary = low\ and\ city = vancouver)$$
$$P(Creditworthiness = bad \mid age = young,\ salary = low\ and\ city = vancouver)$$

Using decision rule of the Naive Bayes-Classifier

$$argmax_{c_j \in C} P(c_j) \cdot \prod_{i=1}^{d} P(x_i \mid c_j)$$

The probability for credit = good when age = young, salary = low and city = Vancouver would be:

$$P(Creditworthiness = good|age = young, salary = low \text{ and } city = vancouver)$$
$$= P(credit = good) \cdot P(Age = young|C = good) \cdot P(Salary = low|C = good) \cdot P(City = Vancouver|C = good)$$
$$= 0.4 * 0.5 * 0.0 * 0.25$$
$$= 0.0$$

The probability for credit = bad when age = young, salary = low and city = Vancouver would be:

$$P(Creditworthiness = bad|age = young, salary = low \text{ and } city = vancouver)$$
$$= P(credit = bad) \cdot P(Age = young|C = bad) \cdot P(Salary = low|C = bad) \cdot P(City = Vancouver|C = bad)$$
$$= 0.6 * 0.5 * 0.5 * 0.5$$
$$= 0.075$$

Since
$P(Creditworthiness = bad|age = young, salary = low \text{ and } city = vancouver) >$
$P(Creditworthiness = good|age = young, salary = low \text{ and } city = vancouver)$
The predict class label (Creditworthiness) using Naïve Bayes classifier for an unseen client with Age = young, Salary = low and City = Vancouver would be **bad.**

1.2. Required: Probability of observing a client with Age = old, Salary = medium and City = Vancouver
Since there is no conditional probability among each parameter, the probability for

$$P(age = old, salary = medium \text{ and } city = vancouver)$$
$$= P(Creditworthiness = bad|age = old, salary = medium \text{ and } city = vancouver)$$
$$+ P(Creditworthiness = good|age = old, salary = medium \text{ and } city = vancouver)$$

$$P(Creditworthiness = good|age = old, salary = medium \text{ and } city = vancouver)$$
$$= P(credit = good) \cdot P(Age = old|C = good) \cdot P(Salary = medium |C = good) \cdot P(City = Vancouver|C = good)$$
$$= 0.4 * 0.5 * 0.75 * 0.25$$
$$= 0.0375$$

$$P(Creditworthiness = bad|age = old, salary = medium \text{ and } city = vancouver)$$
$$= P(credit = bad) \cdot P(Age = old|C = bad) \cdot P(Salary = medium |C = bad) \cdot P(City = Vancouver|C = bad)$$
$$= 0.6 * 0.16 * 0.34 * 0.5$$
$$= 0.01632$$

$$P(age = old, salary = medium \text{ and } city = vancouver) = 0.0375 + 0.01632$$
$$= 0.05382$$

Probability of observing a client with Age = old, Salary = medium and City = Vancouver is **0.05382**

2.

2.1. We would be using self-training (the simplest form of semi-supervised classification) in this situation. We would first build a classifier using the provided labeled dataset (L). Once we have a classifier, we will try to label the entire unlabeled dataset (U). From this newly labeled dataset (P) we will select the tuple (t) with the highest confident label prediction and add it to the set of labeled data (L). We will keep repeating this process using new labeled dataset (L + t), until we are able to label all the data points in U.

2.2. Proposed Algorithm

Input: Labeled dataset (L) , Unlabeled dataset (U)
Output: U' (the class labels for all examples in U)
Algorithm:

Classification (labeled dataset L, unlabeled dataset U):

$U' = \{\}$ *# Initialize empty set U'*
$P = \{\}$ *# Initialize empty set P*

repeat until $U = \emptyset$:

$h = TrainClassifier\ (labeled\ dataset\ L)$ *# h is the classifier trained on*

for every data point (t_i) in unlabeled dataset U:

 # get the labeled data point and confidence for each t_i

$t_i'{}_{(labeled\ datapoint)}, \epsilon_{confidence\ of\ the\ prediction} = Classifier\ (t_i, h)$

 # add this labeled data point and confidence to P

$P = P \cup (t_i'{}_{(labeled\ datapoint)}, \epsilon_{confidence\ of\ the\ prediction})$

 # out of all the newly labeled data P select one data point t_i with highest ϵ_{cop}

$t = t_i'\ in\ P\ having\ the\ maximim\ \epsilon_{confidence\ of\ the\ prediction}\ value$

$P = \{\}$ *# Reinitialize P*
$L = L + t$ *# Add t to our labeled dataset L*
$U = U - t$ *# Remove t from unlabeled dataset*
$U' = U' \cup t$ *# Add the newly labeled data point to U'*

return U' *# U' has all the newly labeled data points*