# CS-760 Machine Learning
# Assignment -4 Solution

## 1. Solution - 1

Entropy of the class is given by,

$$H(Y) = -\sum_{i=1}^{k} p_i log_2 p_i$$

Conditional Entropy

$$H(Y|X = v) = -\sum_{i=1}^{k} Pr(Y = y_i|X = v) log_2 Pr(Y = y_i|X = v)$$

$$H(Y|X) = \sum_{v values of X}^{k} Pr(X = v) H(Y|X = v)$$

Y is a label. X is an attribute or a question. v is an answer to a question.

Information gain : $I(Y;X) = H(Y) - H(Y|X)$

### 1.a. Information gained by knowing whether or not the value of feature C is less than 475

H(class) = $-2/5 log_2(2/5) - 3/5 log_2(3/5) = 0.97095 \approx 0.9710$

**Information gain if we take value of C less/greater than 475**

$H(class|C) = (2/5) H(class|C <= 475) + (3/5) H(class|C > 475)$
$H(class|C) = (2/5)(-1/2 log_2(1/2) - 1/2 log_2(1/2)) + (3/5)(-1/3 log_2(1/3) - 2/3 log_2(2/3))$
$H(class|C) = (2/5)(1) + (3/5)(0.918295)$
$H(class|C) = 0.950977$
$I(class; C) = H(class) - H(class|C) = 0.97095 - 0.950977 = 0.01997 \approx 0.02$
Information gained by knowing whether or not the value of feature C is less than 475 is 0.02.

1.b. Information gained by knowing whether or not the value of features A and B are different.

H(class) = $-2/5 log_2(2/5) - 3/5 log_2(3/5) = 0.97095$

**Information gain if values of A and B are same/different.**
$H(class|AB) = (2/5)H(class|AandBdifferent) + (3/5)H(class|AandBsame)$
$H(class|AB) = (2/5)(-0 - 1log_2(1)) + (3/5)(-1log_2(1) - 0)$
$H(class|AB) = (2/5)(0) + (3/5)(0)$
$H(class|AB) = 0$
$I(class; AB) = H(class) - H(class|AB) = 0.97095 - 0 = 0.97095 \approx 0.971$
Information gained by knowing whether or not the value of features A and B are different is 0.971.

## 2. Solution - 2

**Leave One Out Cross Validation (LOOCV)**

**KNN, with k=1**

k = 1
Instance = 1 Class = Positive
Closest instance is Instance - 2 with Manhattan distance 3 Classified correctly
**Correct**
Instance = 2 Class = Positive
Closest instance is Instance - 3 with Manhattan distance 1 Classified incorrectly
**Incorrect**
Instance = 3 Class = Negative
Closest instance is Instance - 2 with Manhattan distance 1 Classified incorrectly
**Incorrect**
Instance = 4 Class = Positive
Closest instance is Instance - 5 with Manhattan distance 2 Classified incorrectly
**Incorrect**
Instance = 5 Class = Negative
Closest instance is Instance - 6 with Manhattan distance 1 Classified correctly
**Correct**
Instance = 6 Class = Negative
Closest instance is Instance - 5 with Manhattan distance 1 Classified correctly
**Correct**
**Total correctly classified instances are 3**

**KNN, with k=2**
k = 2
Instance = 1 Class = Positive

Closest instance is Instance - 2 with Manhattan distance 3 Classified correctly
Closest instance is Instance - 3 with Manhattan distance 4 Classified incorrectly
**Correct**
Instance = 2 Class = Positive
Closest instance is Instance - 3 with Manhattan distance 1 Classified incorrectly
Closest instance is Instance - 1 with Manhattan distance 3 Classified correctly
**Correct**
Instance = 3 Class = Negative
Closest instance is Instance - 2 with Manhattan distance 1 Classified incorrectly
Closest instance is Instance - 1 with Manhattan distance 4 Classified incorrectly
**Incorrect**
Instance = 4 Class = Positive
Closest instance is Instance - 5 with Manhattan distance 2 Classified incorrectly
Closest instance is Instance - 2 with Manhattan distance 3 Classified correctly
**Correct**
Instance = 5 Class = Negative
Closest instance is Instance - 6 with Manhattan distance 1 Classified correctly
Closest instance is Instance - 4 with Manhattan distance 2 Classified incorrectly
**Incorrect**
Instance = 6 Class = Negative
Closest instance is Instance - 5 with Manhattan distance 1 Classified correctly
Closest instance is Instance - 4 with Manhattan distance 3 Classified incorrectly
**Incorrect**
**Total correctly classified instances are 3**

**KNN, with k=3**
k = 3
Instance = 1 Class = Positive
Closest instance is Instance - 2 with Manhattan distance 3 Classified correctly
Closest instance is Instance - 3 with Manhattan distance 4 Classified incorrectly
Closest instance is Instance - 4 with Manhattan distance 4 Classified correctly
**Correct**
Instance = 2 Class = Positive
Closest instance is Instance - 3 with Manhattan distance 1 Classified incorrectly
Closest instance is Instance - 1 with Manhattan distance 3 Classified correctly
Closest instance is Instance - 4 with Manhattan distance 3 Classified correctly
**Correct**
Instance = 3 Class = Negative
Closest instance is Instance - 2 with Manhattan distance 1 Classified incorrectly
Closest instance is Instance - 1 with Manhattan distance 4 Classified incorrectly
Closest instance is Instance - 4 with Manhattan distance 4 Classified incorrectly
**Incorrect**
Instance = 4 Class = Positive

Closest instance is Instance - 5 with Manhattan distance 2 Classified incorrectly
Closest instance is Instance - 2 with Manhattan distance 3 Classified correctly
Closest instance is Instance - 6 with Manhattan distance 3 Classified incorrectly
**Incorrect**
Instance = 5 Class = Negative
Closest instance is Instance - 6 with Manhattan distance 1 Classified correctly
Closest instance is Instance - 4 with Manhattan distance 2 Classified incorrectly
Closest instance is Instance - 2 with Manhattan distance 5 Classified incorrectly
**Incorrect**
Instance = 6 Class = Negative
Closest instance is Instance - 5 with Manhattan distance 1 Classified correctly
Closest instance is Instance - 4 with Manhattan distance 3 Classified incorrectly
Closest instance is Instance - 2 with Manhattan distance 4 Classified incorrectly
**Incorrect**
**Total correctly classified instances are 2**

After learning the k-nearest neighbor model using the LOOCV we can select either k=1 or k=2 as the both gave the same results and classified the same number of instances correctly. If following the Occam's Razor, k=1 can be opted, else any of k=1 and k=2 can be output as the value of k in k-Nearest Neighbor model after LOOCV.

## 3. Solution - 3

**Sparse Candidate Algorithm**

### 3.a. Compute Mutual information between features.

$$I(X,Y) = \sum_{x,y} P(x,y)log_2 \frac{P(x,y)}{P(x)P(y)}$$

#### 3.a.1. Compute Mutual information between Z and X i.e. I(X,Z)

$P(X=T,Z=T)log_2 \frac{P(X=T,Z=T)}{P(X=T)P(Z=T)} = 0.38log_2 \frac{0.38}{0.50*0.55}$
$P(X=T,Z=F)log_2 \frac{P(X=T,Z=F)}{P(X=T)P(Z=F)} = 0.12log_2 \frac{0.12}{0.50*0.45}$
$P(X=F,Z=T)log_2 \frac{P(X=F,Z=T)}{P(X=F)P(Z=T)} = 0.17log_2 \frac{0.17}{0.50*0.55}$
$P(X=F,Z=F)log_2 \frac{P(X=F,Z=F)}{P(X=F)P(Z=F)} = 0.33log_2 \frac{0.33}{0.50*0.45}$

$$I(X,Z) = 0.132844961809 \approx 0.1328$$

### 3.A.2. COMPUTE MUTUAL INFORMATION BETWEEN Z AND Y I.E. I(Y,Z)

$P(Y=T,Z=T)log_2\frac{P(Y=T,Z=T)}{P(Y=T)P(Z=T)} = 0.45log_2\frac{0.45}{0.50*0.55}$
$P(Y=T,Z=F)log_2\frac{P(Y=T,Z=F)}{P(Y=T)P(Z=F)} = 0.05log_2\frac{0.05}{0.50*0.45}$
$P(Y=F,Z=T)log_2\frac{P(Y=F,Z=T)}{P(Y=F)P(Z=T)} = 0.10log_2\frac{0.10}{0.50*0.55}$
$P(Y=F,Z=F)log_2\frac{P(Y=F,Z=F)}{P(Y=F)P(Z=F)} = 0.40log_2\frac{0.40}{0.50*0.45}$

$$I(Y,Z) = 0.397312609749 \approx 0.3973$$

## 3.b. Which feature should be selected as candidate parent for Z

Y should be selected as the candidate parent for Z, as it has better mutual information between Z than X.

## 3.c. Estimate the parameters of the current Bayes net.

P(X)

| **T** | **F** |
|---|---|
| 0.5 | 0.5 |

P(Y|X)

| **X \ Y** | **T** | **F** |
|---|---|---|
| T | 0.8 | 0.2 |
| F | 0.2 | 0.8 |

P(Z|Y)

| **Y \ Z** | **T** | **F** |
|---|---|---|
| T | 0.9 | 0.1 |
| F | 0.2 | 0.8 |

## 3.d. Kullback-Leibler divergence between the marginal distributions of X and Z as estimated from data and the current network

Kullback-Leibler (KL) divergence provides a distance measure between two distributions, P and Q

$$D_{KL}(P(X)||Q(X) = \sum_x P(x)log\frac{P(x)}{Q(x)}$$

In our case, $P(X) = \hat{P}(X,Z)$ and $Q(X) = P_{net}(X,Z)$
Calculating $\hat{P}(X,Z)$ implies,
$\hat{P}(X=T,Z=T) = 0.38$
$\hat{P}(X=T,Z=F) = 0.12$

$\hat{P}(X = F, Z = T) = 0.17$
$\hat{P}(X = F, Z = F) = 0.33$

Calculating $P_{net}(X, Z)$ implies,

$$P_{net}(X,Z) = \sum_{y} P(X)P(Y|X)P(Z|Y)$$

$P_{net}(X = T, Z = T) = (P(X = T)P(Y = T|X = T)P(Z = T|Y = T)) + (P(X = T)P(Y = F|X = T)P(Z = T|Y = F))$
implies, $P_{net}(X = T, Z = T) = 0.5 \times 0.8 \times 0.9 + 0.5 \times 0.2 \times 0.2 = 0.38$
Similarly,
$P_{net}(X = T, Z = T) = 0.5 \times 0.8 \times 0.1 + 0.5 \times 0.2 \times 0.8 = 0.12$
$P_{net}(X = T, Z = T) = 0.5 \times 0.2 \times 0.9 + 0.5 \times 0.2 \times 0.8 = 0.17$
$P_{net}(X = T, Z = T) = 0.5 \times 0.2 \times 0.1 + 0.5 \times 0.8 \times 0.8 = 0.33$
Now, KL Divergence between the marginal distributions of X and Z as estimated from data and the current network will be,

$$D_{KL}(\hat{P}(X, Z)||P_{net}(X, Z) = \sum_{x,z} \hat{P}(X, Z) log \frac{\hat{P}(X, Z)}{P_{net}(X, Z)}$$

For all the values of X and Z since the values of $\hat{P}(X, Z)$ and $P_{net}(X, Z)$ are same hence the log term will 0. Hence the KL divergence between them is 0.

### 3.e. Should we consider X as a candidate parent of Z?

Since the KL divergence of the distribution is 0 hence X and Z are independent and there is no gain in consider X as a candidate parent for Z.

4.

Instance 1 : (12, 4)
Instance 2 : (3, 18)
Instance 3 : (6, 11)
Instance 4 : (5, 5)

(a) Polynomial Kernel of degree 2.
K(Instance A, Instance B) = (dot(Instance A, Instance B) )^2

| | Instance 1 | Instance 2 | Instance 3 | Instance 4 |
|---|---|---|---|---|
| Instance 1 | 25600 | 11664 | 13456 | 6400 |
| Instance 2 | 11664 | 110889 | 46656 | 11065 |
| Instance 3 | 13456 | 46656 | 24649 | 7225 |
| Instance 4 | 6400 | 11065 | 7225 | 2500 |

(b) Polynomial Kernel of degree up to 2.
K(Instance A, Instance B) = (dot(Instance A, Instance B) + 1)^2

| | Instance 1 | Instance 2 | Instance 3 | Instance 4 |
|---|---|---|---|---|
| Instance 1 | 25921 | 11881 | 13689 | 6561 |
| Instance 2 | 11881 | 111556 | 47089 | 11236 |
| Instance 3 | 13689 | 47089 | 24964 | 7396 |
| Instance 4 | 6561 | 11236 | 7396 | 2601 |

(c) RBF Kernel with $\gamma = 1$.

| | Instance 1 | Instance 2 | Instance 3 | Instance 4 |
|---|---|---|---|---|
| Instance 1 | 1 | ~0 | ~0 | ~0 |
| Instance 2 | ~0 | 1 | ~0 | ~0 |
| Instance 3 | ~0 | ~0 | 1 | ~0 |
| Instance 4 | ~0 | ~0 | ~0 | 1 |

5.

The VC-dimension of the hypothesis space is 2 (For any set of 3 instances, there are labelings for which we cannot find a consistent hypothesis). Thus, the sample complexity grows polynomially in $\frac{1}{\varepsilon}$ and $1/\delta$

$$m = \frac{1}{\varepsilon}\ (4\log\frac{2}{\delta} + \ 8\,Vc - \dim(H)\log\frac{13}{\varepsilon})$$

We can specify a polynomial time algorithm for finding consistent hypothesis:

1. Sort training instances by distance from origin.
2. Set r to be < distance to first pos in the sorted list.
3. Set r to be > distance to last pos in the sorted list.