Homework1

Problem 1:

a. The algorithm used for prototype selection is inspired by the CNN (condensed nearest neighbor) algorithm.
CNN: Let TR be the training set and PS be the prototype set. We initialize the prototype set to be an empty set. Each instance *(i)* in TR is classified using only the instances in PS. If the instance *(i)* is misclassified, it is added to PS. This continues till there are no more elements to be classified in TR.

b. Pseudocode:

```
// Classification using CNN (Condensed nearest neighbor)
// Method: get_cnn_prototype_set
// Input: training_set(images, labels), M (size of the prototype_set)
// Output: p_set(images, labels) of size M

Initialize empty prototype set p_set(images, labels) and r_set(images, labels)
// r_set contains samples which are not selected in the p_set
Randomly shuffle the training_set {(images, labels)}

while(training_set) && sizeof(p_set) < M:
  sample_set(image, label) = remove a sample from the training_set
  // calculate_nearest_neighbor will return the label of the nearest neighbor in the
  training set

  if( sample_set[label] != calculate_nearest_neighbor(sample_set[image],
  proto_training_set)):
           Add sample_set to p_set
  Else:
           Add sample_set to r_set

if(sizeof(p_set) < M)
  Randomly select max((M - sizeof(p_set)), sizeof(r_set)) samples from r_set and add them
  to the p_set

 return p_set

// Note:
// the images are classified using 1NN
// Distance measure used = Eucledian distance = ||x-y||₂
```

c. Comparing error rates for classification with random prototype vs. CNN prototype of size M

Error Rates →	Random Prototype Selection			CNN Prototype Selection		
	Run:1	Run:2	Run:3	Run:1	Run:2	Run:3
M = 1000	0.114	0.1112	0.1151	0.1115	0.1146	0.1086
M = 5000	0.0659	0.0664	0.0635	0.0615	0.0639	0.0626
M = 10000	0.0509	0.0522	0.0528	0.0507	0.0513	0.0517

M	Random Prototype Selection		CNN Prototype Selection		Average Improvement
	Mean	Variance	Mean	Variance	
1000	0.1134	$2.67 * 10^{-6}$	0.1116	$6 * 10^{-6}$	1.6%
5000	0.0653	$1.6 * 10^{-6}$	0.0627	$9.9 * 10^{-7}$	3.98%
10000	0.0520	$6.3 * 10^{-7}$	0.0512	$1.7 * 10^{-7}$	1.54%

Problem 2:

a. The Bayes-optimal classifier: $h^*(x) = \begin{Bmatrix} 1 & if\ \eta > 1/2 \\ 0 & otherwise \end{Bmatrix}$

$$h^*(x) = \begin{Bmatrix} 0 & if\ x < -0.5 \\ 1 & if -0.5 \le x \le 0.5 \\ 0 & if\ x > 0.5 \end{Bmatrix}$$

Optimal risk: $R^* = R(h^*) = E_x \min(\eta(X), 1 - \eta(X))$

$R^* = \int |x| \min(\eta(X), 1 - \eta(X))$

$= \int_{-1}^{-0.5}(-x)0.2 + \int_{-0.5}^{0}(-x)0.2 + \int_{0}^{0.5} x * 0.2 + \int_{0.5}^{1} x * 0.4$

// solving the integrals

$= 0.1 * 0.75 + 0.1 * 0.25 + 0.1 * 0.25 + 0.2 * 0.75 = 0.275$

b. From the training set the decision boundary is as follows:

$$h(x) = \begin{Bmatrix} 0 & if\ x < -0.6 \\ 1 & if -0.6 \le x \le 0.5 \\ 0 & if\ x > 0.5 \end{Bmatrix}$$

True error rate of the classifier: $= P(h(x)\ != Y)$

$= P(h(x) = 0, Y = 1) + P(h(x) = 1, Y = 0)$

$= \int_{-1}^{1} P(x, h(x) = 0, Y = 1) + P(x, h(x) = 1, Y = 0)\ .\ dx$

// Using marginalization and conditional independance

$= \int_{-1}^{1} P(x). P(h(x) = 0 |\ x). P(Y = 1|x) + P(x). P(h(x) = 1 |\ x). P(Y = 0|x). dx$

$= \int_{-1}^{1} |x|. P(h(x) = 0 |\ x). P(Y = 1|x) + |x|. P(h(x) = 1 |\ x). P(Y = 0|x). dx$

$= A + B$

$where\ A = \int_{-1}^{1} |x|. P(h(x) = 0 |\ x). P(Y = 1|x)\ \ dx$

$B = \int_{-1}^{1} |x|. P(h(x) = 1 |\ x). P(Y = 0|x). dx$

$A = \int_{-1}^{1} |x|. P(h(x) = 0 |\ x). P(Y = 1|x)\ \ dx$

h(x) = 0 only when *x < -0.6 and x > 0.5*

$= \int_{-1}^{-0.6} -x. P(Y = 1|x)\ \ dx + \int_{0.5}^{1} x. P(Y = 1|x)\ \ dx$

$= \int_{-1}^{-0.6} -x * 0.2\ \ dx + \int_{0.5}^{1} x * 0.4\ \ dx$

$= 0.064 + 0.15$

$A = 0.214$

$B = \int_{-1}^{1} |x|.P(h(x) = 1|\, x).P(Y = 0|x).dx$

$h(x) = 1$ only when $-0.6 \le x \le 0.5$

$= \int_{-0.6}^{-0.5} -x.P(Y = 0|x)\ dx + \int_{-0.5}^{0} x.P(Y = 0|x) + \int_{0}^{0.5} x.P(Y = 0|x)\ dx$

$= \int_{-0.6}^{-0.5} -x * 0.8\ dx + \int_{-0.5}^{0} -x * 0.2 + \int_{0}^{0.5} x * 0.2\ dx$

$= 0.044 + 0.025 + 0.025$

$B = 0.094$

$\boldsymbol{True\ error = 0.308}$

c. $R(h) = c_{01}P(Y = 0, h(x) = 1) + c_{10}P(Y = 1, h(x) = 0)$

$= R_1 + R_0\ where\ R_1 = c_{01}P(Y = 0, h(x) = 1)\ and\ R_0 = c_{10}P(Y = 1, h(x) = 0)$

At any given x, $h(x) = 0\ if R_1 > R_0\ and\ h(x) = 1\ otherwise$

$h(x) = 0\ if\ c_{01}P(Y = 0\ |x) > c_{10}P(Y = 1\ |x)\ and\ h(x) = 1\ otherwise$

$h(x) = 0\ if\ c_{01}(1 - \eta(x)) > c_{10}(\eta(x))\ and\ h(x) = 1\ otherwise$

	$c_{01}(1 - \eta(x))$	$c_{10}(\eta(x))$	$h(x)$
x < 0.5	0.8	0.1 * 0.2 = 0.02	0
-0.5<= x <= 0.5	0.2	0.1*0.8 = 0.08	0
x>0.5	0.6	0.1*0.4 = 0.04	0

The decision boundary would be when $c_{01}(1 - \eta(x)) = c_{10}(\eta(x))$

d. $R(h) = c_{01}P(Y = 0, h(x) = 1) + c_{10}P(Y = 1, h(x) = 0)$

From part c,

At any given x, $h(x) = 0\ if R_1 > R_0\ and\ h(x) = 1\ otherwise$

$h(x) = 0\ if\ c_{01}P(Y = 0\ |x) > c_{10}P(Y = 1\ |x)\ and\ h(x) = 1\ otherwise$

$h(x) = 0\ if\ c_{01}(1 - \eta(x)) > c_{10}(\eta(x))\ and\ h(x) = 1\ otherwise$

$\boldsymbol{h(x) = 0\ if\ \eta(x) < \frac{c_{01}}{c_{01} + c_{10}}\ and\ h(x) = 1\ otherwise}$

Problem 3:

a. L_1 distance : $\boldsymbol{d(x, y) = ||x - y||_1 = \sum_{i=1}^{m} |x_i - y_i|}$

Hence, $d(x, y) \ge 0$

$d(x, y) = 0\ iff\ |x_i - y_i| = 0\ \forall i, ie\ when\ x = y$

$d(x, y) = ||x - y||_1 = \sum_{i=1}^{m} |x_i - y_i| = \sum_{i=1}^{m} |y_i - x_i| = ||y - x||_1 = d(y, x)$

Triangle Inequality:

$|x_i - z_i| \le |x_i - y_i| + |y_i - z_i|\ \forall y_i\ (Using\ triangle\ inequality)$

// Taking summation over all i

$\sum_{i=1}^{m} |x_i - z_i| \le \sum_{i=1}^{m} |x_i - y_i| + \sum_{i=1}^{m} |y_i - z_i|$

$d(x, z) \le d(x, y) + d(y, z)$

<u>**Therefore, it is a metric**</u>

b. $d_1 + d_2$ where $d_1\ d_2$ are both metrics

Proof:
$d(x, y) = d_1(x, y) + d_2(x, y)$

1.
$d_1(x, y) \geq 0$ and $d_2(x, y) \geq 0$
$d(x, y) = d_1(x, y) + d_2(x, y) \geq 0$

2.
From 1, $d(x, y) = 0$ iff $d_1(x, y) == 0$ and $d_2(x, y) == 0$
Which implies $d(x, y) = 0$ iff $x == y$

3.
$d(x, y) = d_1(x, y) + d_2(x, y) = d_1(y, x) + d_2(y, x) = d(y, x)$

4.
Triangle inequality:
$d_1(x, z) \leq d_1(x, y) + d_1(y, z)$ and $d_2(x, z) < d_2(x, y) + d_2(y, z)$
Adding up, $d_1(x, z) + d_2(x, z) \leq d_1(x, y) + d_1(y, z) + d_2(x, y) + d_2(y, z)$
$d(x, z) \leq d(x, y) + d(y, z)$

Therefore, it is a metric

c. **d(x, y) = # of positions on which x and y differ**
$d(x, y) \geq 0$ $(since\ we\ are\ counting\ the\ number\ of\ positions)$
$d(x, y) = 0$ $iff\ x\ and\ y\ differ\ by\ 0\ positions, ie. x = y$
$d(x, y) = d(y, x)$ $(since\ x\ and\ y\ differe\ by\ same\ number\ of\ positions\ as\ y\ and\ x$
We can turn x to y by changing atmost d(x, y) characters and turn y to z by changing atmost d(y, z) characters. So to turn x into z will change no more than d(x, y) + d(y, z) characters.
Hence,
$d(x, z) \leq d(x, y) + d(y, z)$ S

Therefore, it is a metric

d. $\boldsymbol{d(x, y) = \sum_{i=1}^{m} (x_i - y_i)^2}$
Proof by contradiction:
According to the triangle rule of inequality: $d(x, z) < d(x, y) + d(y, z)$
Let m = 1, $x_1 = 4$, $y_1 = 2$, $z_1 = 1$
Substituting values in the triangle inequality,
$LHS = 3^2 = 9$, $RHS = 2^2 + 1^2 = 5$
$LHS > RHS$

Therefore, it is not a metric

e. $\boldsymbol{d(p, q) = K(p, q) = \sum_{i=1}^{m} p_i * \log \frac{p_i}{q_i}}$
Proof by contradiction:
Given: $X = \{p \in R^m: p_i \geq 0, \sum_i p_i = 1\}$, Let m = 2 and $p_1 = 0.5$ and $p_2 = 0.5$
$Y = \{q \in R^m: q_i \geq 0, \sum_i q_i = 1\}$, Let m = 2 and $q_1 = 0.2$ and $q_2 = 0.8$
$K(p, q) = \sum_{i=1}^{m} p_i * \log \frac{p_i}{q_i} = 0.5 * \log \frac{0.5}{0.2} + 0.5 * \log \frac{0.5}{0.8} = 0.2231$

$K(q,p) = \sum_{i=1}^{m} q_i * \log\frac{q_i}{p_i} = 0.2 * \log\frac{0.2}{0.5} + 0.8 * \log\frac{0.8}{0.5} = 0.1927$
$K(q,p) \,! = K(p,q)$

<u>Therefore, it is not a metric</u>