CS 260 : Machine Learning Algorithm
Homework #1

Ronak Sumbaly
UID - 604591897

October 6, 2015

Problem 1 - Sequence of Coin Flips

(a) **Canonical Setup**

Let X be a random variable that counts the number of trials until the first heads shows up, where the probability of heads is p.

Data

The sample space Ω of X consisting of all sequences over the alphabets $H = Heads, T = Tails$ and where each sample event ends with H and contains no other H is given by

$$\Omega = \{H, TH, TTH, ...\} \tag{1}$$

Considering the subspace ω where $\omega \in \Omega$ then
$P(X = 1)$ is the probability that it takes one coin flip to obtain a heads on the first trial.

$$\mathbb{P}[X = 1] = p \qquad \text{and} \qquad \omega = \{H\} \tag{2}$$

$P(X = 2)$ is the probability that it takes two coin flips until a heads appears. Thus the first trial must come up a tails and the second comes up heads. The two coin flips are independent events, so we multiply their probability. We have $(1 - p)$ for the first trial and p for the second. We conclude

$$\mathbb{P}[X = 2] = (1 - p) \cdot p \qquad \text{and} \qquad \omega = \{TH\} \tag{3}$$

Similarly, $P(Y = 3)$ requires two flips of tails and then a flip of heads and so we have corresponding probabilities of $(1 - p)$, $(1 - p)$, and p. The events are independent so we multiple to obtain

$$\mathbb{P}[X = 3] = (1 - p)^2 \cdot p \qquad \text{and} \qquad \omega = \{TTH\} \tag{4}$$

We can therefore conclude that X follows a Geometric Distribution with parameter p and write
$X \sim Geo(p)$ and generally, for any sequence $\omega \in \Omega$ and for $X = k$

$$\mathbb{P}[X = k] = (1 - p)^{k-1} \cdot p \tag{5}$$

(b) **Canonical Setup**

The probability function for the random variable X that counts the number of trials until the first heads shows up, where the probability of heads is p is given by

$$\mathbb{P}[X = k] = (1-p)^{k-1} \cdot p \tag{6}$$

Prove

To show for any $x_0 \geq 1$ the probability is

$$\mathbb{P}[X \geq x_0] = (1-p)^{x_0-1}$$

Proof

Consider the expression $X \geq x_0$, we want to calculate the probability that its take x_0 or more than x_0 flips before we get a heads. This means it could take x_0 flips, $x_0 + 1$ flips etc. However, in each of these cases, we have that the first $x_0 - 1$ flips were are all tails. In fact, the event $X \geq x_0$ is equivalent to the event the first $x_0 - 1$ events were all tails. Therefore,

$$\mathbb{P}[X \geq x_0] = P[\text{first } x_0 - 1 \text{ flips were tails}] \tag{7}$$

Remark 1: From (6) we know that the probability of $x_0 - 1$ events to occur with only tails is $(1-p)^{x_0-1}$.

Alternatively, we can also say that $\mathbb{P}[X \geq x_0]$ is equal to the probability of $X \neq 1, 2, 3 \ldots, x_0 - 1$

$$\mathbb{P}[X \geq x_0] = \mathbb{P}[X \neq 1, X \neq 2, X \neq 3, \ldots, X \neq x_0 - 1] \tag{8}$$

$$= \mathbb{P}[X \neq 1] \cdot \mathbb{P}[X \neq 2 | X \neq 1] \cdot \mathbb{P}[X \neq 3 | X \neq 2] \ldots \mathbb{P}[X \neq x_0 - 2 | X \neq x_0 - 1] \text{ using Bayes Rule} \tag{9}$$

$$= (1-p) \cdot \frac{(1-p)^2}{(1-p)^1} \cdots \frac{(1-p)^{x_0-1}}{(1-p)^{x_0-2}} = (1-p)^{x_0-1} \tag{10}$$

Hence from Remark 1 and (11) we can conclude,

$$P[X \geq x_0] = (1-p)^{x_0-1} \tag{11}$$

(c) **Canonical Setup**

We have been given that p is uniformly distributed in $[0, 1]$ interval. Also, assume that the first coin flip equals to heads.

Data

To compute the probability of $p > 1/2$ given that the first coin flip equals to heads is given by

$$P[p > 1/2|X = 1] \tag{12}$$

Using Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{13}$$

(12) can be written as

$$P[p > 1/2|X = 1] = \frac{P[X = 1|p > 1/2] \cdot P[p > 1/2]}{P(X = 1)}$$

1. $P(X = 1) = p$ ... Since we have assumed that the first coin flip equals to heads, we can say that the probability of this occurring is 1
2. $P[p > 1/2]$ ... Since p follows uniform distribution, using probability function of uniform distribution in the range $[0, 1]$ we get

$$\begin{aligned} P[p > x]\text{in range}[c, d] &= \frac{x - c}{d - c} \\ &= \frac{1/2 - 0}{1 - 0} \\ &= 1/2 \end{aligned}$$

3. $P[X = 1|p > 1/2] = p$ ... The probability of the coin giving heads on the first try given that the $P(Heads) > 1/2$ is basically equal to the probability of heads appearing on the first try.

Combining the results of $1, 2$ and 3 we get,

$$P[p > 1/2|X = 1] = 1/2 \tag{14}$$

As seen from (14) we can conclude that the events are independent of each other and there is no increase or decrease in $P[p > 1/2]$

Problem 2 - Convex Functions and Information Theory

(a) **Prove**

Show that the function $f(x) = |x| + \exp(x)$ is convex

Proof

A function $f(x)$ is convex if, for any two points x_1, x_2 in its domain and for any t in the interval [0, 1], the following is true,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \tag{15}$$

A function $f(x)$ is also convex when,

$$\frac{d^2 f(x)}{dx^2} \geq 0 \qquad \forall x \in Domain \tag{16}$$

Geometrically speaking, if you select any two points on the graph of $f(x)$ and connect them with a line segment, then that line segment lies entirely above the graph of $f(x)$.

Theorem: If $f(x)$ is of the form, $f(x) = g(x) + h(x)$, $f(x)$ is said to be convex if both $g(x), h(x)$ are convex.

Therefore, in order to prove the $|x| + \exp(x)$ to be convex. Consider $g(x) = |x|$ and $h(x) = \exp(x)$

Proof 1. Proving $g(x) = |x|$ is convex.

Using definition (14) $\forall x_1, x_2 \in R$ and $t \in [0, 1]$ we need to prove,

$$|t \cdot x_1 + (1-t) \cdot x_2| \leq |t \cdot x_1| + |(1-t) \cdot x_2| \tag{17}$$

Case 1. $\forall x_1, x_2 \geq 0$ and $\forall x_1, x_2 < 0$

$$|t \cdot x_1 + (1-t) \cdot x_2| = |tx_1| + |(1-t) \cdot x_2| \tag{18}$$

Case 2. $\forall x_2 < 0$ and $x_1 > 0$
Since, the term $(1-t) \cdot x_2$ on the LHS would be less than $|(1-t) \cdot x_2|$ on the RHS we get,

$$|t \cdot x_1 + (1-t) \cdot x_2| < |tx_1| + |(1-t) \cdot x_2| \tag{19}$$

Using the formula $|A+B| \leq |A|+|B|$ and from (17) and (18) we can conclude, $g(x) = |x|$ is convex.

Proof 2. Proving $h(x) = \exp(x)$ is convex.

Using definition (15) of convexity we have,

$$\frac{de^x}{dx} = e^x > 0 \qquad \forall x \in R \tag{20}$$

Thus $h(x) = \exp(x)$ is convex.

Combining $Proof$ 1 and $Proof$ 2 we can conclude,

$$f(x) = |x| + \exp(x) \text{ is convex}$$

(b) **Canonical Setup**

Let X be a random variable distributed according to a k-class multi-nominal distribution with class probabilities $p_1, p_2, \ldots, p_k$ such that $\sum_{i=1}^{k} p_i = 1$.

Data

To find values of $p_i, i = 1, \ldots, k$ such that *entropy* of X is maximized.

The entropy H for a random variable X with possible values $x_1, x_2, \ldots, x_n$ and probability mass function $P(X)$ is given by

$$H(X) = E[I(X)] = E[-ln(P(X))] \tag{21}$$

where E is the expected value operator and I is the information content of X.

The entropy can also be written as,

$$H(X) = \sum_i P(x_i) I(x_i) = -\sum_i P(x_i) \log_b P(x_i) \tag{22}$$

In order to maximize (21) we apply the strategy of *Lagrange multiplier*. To optimize a function using *Lagrange multiplier* we consider,

$$\text{maximize } f(x, y) \text{ subject to } g(x, y) = 0$$

To maximize $f(x)$ we study the Lagrange function defined by

$$L(x, y, \lambda) = f(x, y) + \lambda \cdot g(x, y)$$

Considering $f(x, y) = H(X)$ and $g(x, y) = \sum_{i=1}^{k} p_i - 1$ we get,

$$L = H + \lambda(\sum^{i} P_i - 1) = -\sum^{i} P_i \log_b P_i + \lambda(\sum^{i} P_i - 1) \tag{23}$$

Solving (22) by equating $\frac{dL}{dp_i} = 0$ we get,

$$\frac{\partial L}{\partial p_i} = -\frac{\partial \sum^{i} P_i \log_b P_i}{\partial p_i} + \frac{\partial \lambda(\sum^{i} P_i - 1)}{\partial p_i} = -1 - \log P_i + \lambda = 0 \tag{24}$$

$$\log P_i = \lambda - 1$$

$$P_i = e^{\lambda - 1} \tag{25}$$

(24) states that $\forall p \in P_i$ all p have the same value, applying this to the constraint $\sum_{i=1}^{k} p_i = 1$ we get,

$$k \cdot p = 1 \tag{26}$$

$$\text{since } p_1 = p_2 = \cdots = p_K$$

Therefore in order to maximize entropy for the random variable X the probability value should be

$$p_1 = p_2 = p_3 = \ldots p_k = 1/k \tag{27}$$

Problem 3 - Linear Algebra

(a) **Canonical Setup**

The covariance matrix Σ of an random vector X is defined as,

$$\Sigma = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^T] \tag{28}$$

where $\mathbb{E}\mathbf{X}$ is expectation of $\mathbf{X}$

Prove

Show that Σ is $positive - semidefinite$

Proof

Def: A symmetric matrix $\mathbf{M}$ is said to be $positive - semidefinite$ if $\forall$ non-zero column vector z

$$z \cdot M \cdot z^T \geq 0 \tag{29}$$

We know $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ and (i, j) value in the covariance matrix Σ is defined as

$\Sigma_{ij} = \mathbb{E}[(\mathbf{X_i} - \mathbb{E}\mathbf{X_i})(\mathbf{X_j} - \mathbb{E}\mathbf{X_j})]$

Hence the covariance matrix can be represented as,

$$\Sigma = \begin{bmatrix} \mathbb{E}[(\mathbf{X_1} - \mathbb{E}\mathbf{X_1})(\mathbf{X_1} - \mathbb{E}\mathbf{X_1})] & \mathbb{E}[(\mathbf{X_1} - \mathbb{E}\mathbf{X_1})(\mathbf{X_2} - \mathbb{E}\mathbf{X_2})] & \dots & \mathbb{E}[(\mathbf{X_1} - \mathbb{E}\mathbf{X_1})(\mathbf{X_n} - \mathbb{E}\mathbf{X_n})] \\ \mathbb{E}[(\mathbf{X_2} - \mathbb{E}\mathbf{X_2})(\mathbf{X_1} - \mathbb{E}\mathbf{X_1})] & \mathbb{E}[(\mathbf{X_2} - \mathbb{E}\mathbf{X_2})(\mathbf{X_2} - \mathbb{E}\mathbf{X_2})] & \dots & \mathbb{E}[(\mathbf{X_2} - \mathbb{E}\mathbf{X_2})(\mathbf{X_n} - \mathbb{E}\mathbf{X_n})] \\ \vdots & \vdots & \dots & \vdots \\ \mathbb{E}[(\mathbf{X_n} - \mathbb{E}\mathbf{X_n})(\mathbf{X_1} - \mathbb{E}\mathbf{X_1})] & \mathbb{E}[(\mathbf{X_n} - \mathbb{E}\mathbf{X_n})(\mathbf{X_n} - \mathbb{E}\mathbf{X_2})] & \dots & \mathbb{E}[(\mathbf{X_n} - \mathbb{E}\mathbf{X_n})(\mathbf{X_n} - \mathbb{E}\mathbf{X_n})] \end{bmatrix} \tag{30}$$

It can be seen that $\Sigma = \Sigma^T$, hence we can conclude that, $\Sigma = Symmetric\ Matrix$

Considering the definition (28) for $positive{-}semidefinite$ we need to prove, $\forall z$ where z is a non-negative vector

$$\mathbf{z}^T\Sigma\mathbf{z} > 0 = E[\mathbf{z}^T(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^T\mathbf{z}] \tag{31}$$

$$= E[((\mathbf{X} - \mathbb{E}\mathbf{X})\mathbf{z})^T((\mathbf{X} - \mathbb{E}\mathbf{X})\mathbf{z})] \tag{32}$$

$$= ||((\mathbf{X} - \mathbb{E}\mathbf{X})\mathbf{z})||^2 \geq 0 \tag{33}$$

since $\mathbf{z}^T(\mathbf{X} - \mathbb{E}\mathbf{X}) = (\mathbf{X} - \mathbb{E}\mathbf{X})^T\mathbf{z}$

Thus from (32) we can conclude that Σ is positive-semidefinite

(b) **Canonical Setup**

Let $\mathbf{A}$ and $\mathbf{B}$ be two $\mathbb{R}^{DxD}$ symmetric matrix. Suppose $\mathbf{A}$ and $\mathbf{B}$ have the same set of *eigenvectors* $u_1, u_2, \ldots, u_D$ with corresponding *eigenvalues* $\alpha_1, \alpha_2, \ldots, \alpha_D$ for $\mathbf{A}$ and $\beta_1, \beta_2, \ldots, \beta_D$ for $\mathbf{B}$.

Data

By definition, the *eigenvectors* and *eigenvalues* of the symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ is given by,

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \ldots & A_{1D} \\ A_{21} & A_{22} & A_{23} & \ldots & A_{2D} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ A_{D1} & A_{D2} & A_{D3} & \ldots & A_{DD} \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_D \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{bmatrix} \tag{34}$$

Similarly for matrix $\mathbf{B}$,

$$\begin{bmatrix} B_{11} & B_{12} & B_{13} & \ldots & B_{1D} \\ B_{21} & B_{22} & B_{23} & \ldots & B_{2D} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ B_{D1} & B_{D2} & B_{D3} & \ldots & B_{DD} \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_D \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{bmatrix} \tag{35}$$

1. C=A+B
   Since $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvectors: $u_1, u_2, u_3, \ldots, u_D$ the eigenvalues for the sum of both the symmetric matrices will be the sum of the eigenvalues of the individual matrices. Shown as follows,
   $$C = \mathbf{A} + \mathbf{B} \implies Cu_D = \mathbf{A}u_D + \mathbf{B}u_D$$
   from (33) and (34) we know $mathbfAu_D = \alpha_D$ and $\mathbf{B}u_D = \beta_D$. Thus,
   $$C = \mathbf{A} + \mathbf{B} : Eigenvalues = \alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots, \alpha_D + \beta_D \tag{36}$$

2. D=A-B
   Since $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvectors: $u_1, u_2, u_3, \ldots, u_D$ the eigenvalues for the difference of both the symmetric matrices will be the difference of the eigenvalues of the individual matrices. Shown as follows,
   $$D = \mathbf{A} - \mathbf{B} \implies Du_D = \mathbf{A}u_D - \mathbf{B}u_D$$
   from (33) and (34) we know $mathbfAu_D = \alpha_D$ and $\mathbf{B}u_D = \beta_D$. Thus,
   $$D = \mathbf{A} - \mathbf{B} : Eigenvalues = \alpha_1 - \beta_1, \alpha\_\beta_2, \ldots, \alpha_D - \beta_D \tag{37}$$

3. E=AB Since $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvectors: $u_1, u_2, u_3, \ldots, u_D$ the eigenvalues for the product of both the symmetric matrices will be the product of the eigenvalues of the individual matrices. Shown as follows,
   $$E = \mathbf{AB} \implies Eu_D = \mathbf{AB}u_D \implies Eu_D = \mathbf{A}u_D\beta_D \implies Eu_D = \alpha_D\beta_D u_D$$
   Thus the eigenvalues for,
   $$E = \mathbf{AB} : Eigenvalues = \alpha_1 \cdot \beta_1, \alpha_2 \cdot \beta_2, \ldots, \alpha_D \cdot \beta_D \tag{38}$$

4. $E = A^{-1}B$ The eigenvalues for the inverse of a symmetric matrix is the inverse of the eigenvalues of the symmetric matrix. Shown as follows,

$$\mathbf{A}u = \alpha_D u \Longrightarrow \mathbf{A}^{-1}\mathbf{A}u = \alpha_D \mathbf{A}^{-}1\ u \Longrightarrow \mathbf{A}^{-1}u = \frac{1}{\alpha_D}\ u \tag{39}$$

since $\mathbf{A}^{-1}\mathbf{A} = I$

Thus the eigenvalues by (37) and (38) for,

$$F = \mathbf{A}^{-1}\mathbf{B} : Eigenvalues = \beta_1/\alpha_1, \beta_2/\alpha_2, \beta_3/\alpha_3, \ldots, \beta_D/\alpha_D \tag{40}$$

Problem 4 - KNN Classification in MATLAB

k Nearest Neighbor Accuracy

kNN Classification Accuracy			
k	Training	Testing	Validation
1	77.7895	79.4344	75.5784
3	85.7895	88.6889	83.8046
5	88.9474	92.0308	85.8612
7	90.5263	90.7455	86.8895
9	90.3158	91.5167	87.4036
11	90.3158	89.7172	88.9460
13	89.4737	88.9460	87.4036
15	89.2632	87.6607	85.6041
17	87.8947	87.6607	85.6041
19	86.7368	87.4036	85.6041
21	86.3158	86.8895	84.3188
23	85.6842	86.8895	83.8046

Observations

- The knn_classify.m function was ran on the car_train.data as the training data, car_valid.m as the validation data and car_test.m as the testing data with k = 1, 3, 5, . . . , 23
- The highest validation accuracy was found for $k = 11$ as shown in the above table.
- The test accuracy for the corresponding k value $k = 11$ is 89.7172

Decision Boundary

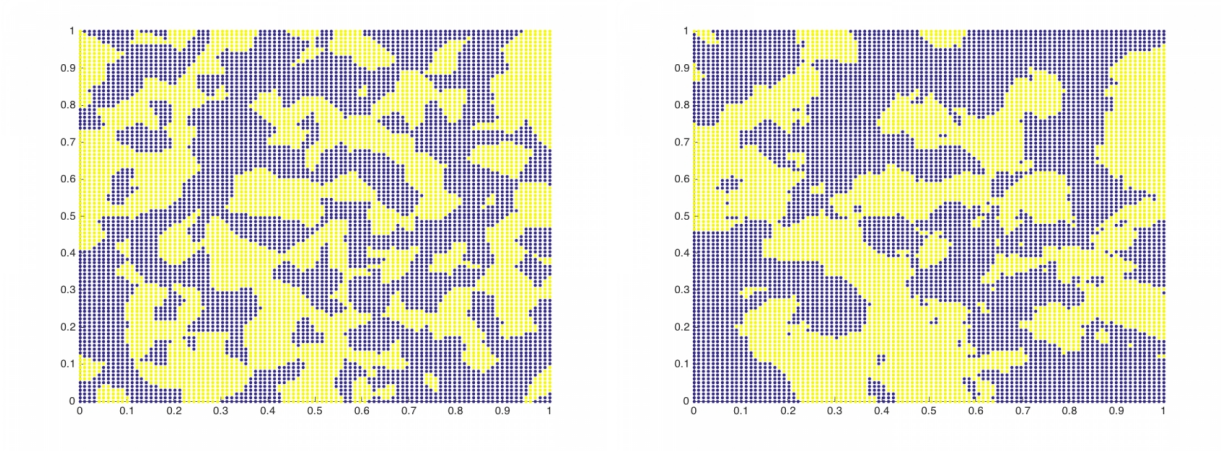

Figure 1: Decision Boundary (Class 1:Yellow, -1:Blue): k=1(Left), k=5(Right)

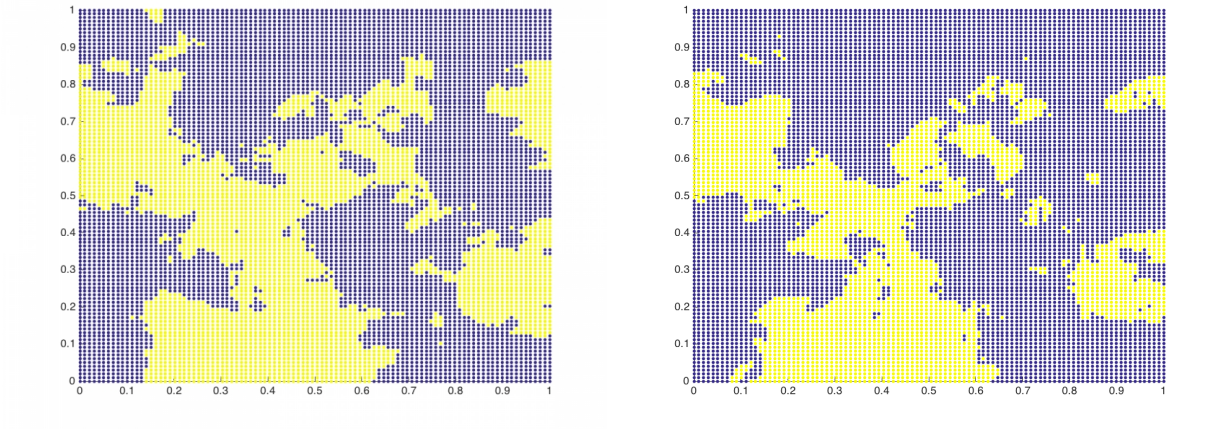

Figure 2: Decision Boundary (Class 1:Yellow, -1:Blue): k=15(Left), k=20(Right)

Observations

- The decision boundary was constructed using 10000 randomly generated points.
- When $k = 1$, the decision boundary is rough with both classes $1, -1$ scattered over the graph.
- It can be seen that as the value of k gradually increases the decision boundary becomes smoother and smoother. When $k = 5$ the majority of classes are in yellow 1 and classification of the points has become smoother as compared to the previous value.
- When $k = 15, 20$ we can see the decision boundary more smooth than the previous values. Increasing the value of k may lead to $overfitting$ in which the entire decision boundary would show a particular class only.

Hence it can be seen that as k value increase the decision boundary becomes smoother.