HW 2

Bryan Yekelchik

March 30, 2022

1 Exercise 1: Decision Tree

Solution:

Entropy of set: $\frac{-5}{9}\log_2\frac{5}{9}+\frac{-5}{9}\log_2\frac{4}{9}=.99$

$IG(All|A)=Entropy(all)-H(all|A)$
$Entropy(A+)=\frac{-1}{6}\log_2\frac{1}{6}+\frac{-5}{6}\log_2\frac{5}{6}$
$Entropy(A-)=0$
$H(All|A)=\frac{-6}{9}*Entropy(A+)+\frac{-3}{9}\log_2 Entropy(A-)=.56$

$IG(All|S)=Entropy(all)-H(all|S)$
$Entropy(S+)=\frac{-3}{5}\log_2\frac{3}{5}+\frac{-2}{5}\log_2\frac{2}{5}=.97$
$Entropy(S-)=\frac{-2}{4}\log_2\frac{2}{4}+\frac{-2}{4}\log_2\frac{2}{4}=1$
$H(All|S)=\frac{-5}{9}*Entropy(S+)+\frac{-5}{9}\log_2 Entropy(S-)=.98$

Because $IG(All|S)>IG(All|A)$, choose A for first split

2 Exercise 2: KNN

X	1.5	3.0	4.4	4.7	4.9	5.1	5.4	5.7	7.5	10
Y	-	-	+	+	+	-	-	+	-	-
D(4.5)	3	1.5	.1	.2	.4	.6	.9	1.5	3	6.5
w(4.5)	.33	.66	10	5	2.5	1.66	1.11	.33	.33	1.5

2.1 a

Classify $x=4.5$ according to its 1,3,5,and 9-nearest neighbors using majority vote.

$$MajorityVoting: y^{'}=argmax_v\sum_{(x_i,y_i)\in D}I(v=y_i)$$

where v is a class label, y_i is the class label for one of the nearest neighbors, and $I(\cdot)$ is an indicator function that returns the value 1 if its argument is true and 0 otherwise.
Solution:

1. $k = 1 \rightarrow +$
2. $k = 3 \rightarrow +$
3. $k = 5 \rightarrow +$ 2 are $(-)$ and 3 are $(+)$
4. $k = 9 \rightarrow -$ 5 are $(-)$ and 4 are $(+)$

2.2 b

Classify $x = 4.5$ according to its 1,3,5,and 9-nearest neighbors using distance-weighted approach.

$$Distance - weighted : y^{'} = argmax_v \sum_{(x_i,y_i)\in D} w_i I(v = y_i)$$

where w_i is the Euclidean Distance

Solution:

1. $k = 1 \rightarrow +.\ 10(1) \rightarrow +$
2. $k = 3 \rightarrow +\ .\ 10(1) + 5(1) + 2.5(1) \rightarrow +$
3. $k = 5 \rightarrow +.\ +1.66(-1) + 1.11(-1) \rightarrow +$
4. $k = 9 \rightarrow +.\ 10(1) + 5(1) + 2.5(1) + 1.66(-1) + 1.11(-1) - .33(3) - .66$

2.3 c

Explain why the distance-weighted voting approach can reduce the impact of K for KNN classifier, compared to the majority vote approach.

Solution: If heavy weight is placed on points closer to the point trying to be classified, then points farthest away will be meaningless. Hence, increasing K can have a reduced effect because the additional nearest points will have lower weight and hence minimal effect on the classification outcome.

3 Exercise 3: Bayes Classifier

Data table:

Record	A	B	C	Class
1	0	0	1	-
2	1	0	1	+
3	0	1	0	-
4	1	0	0	-
5	1	0	1	+
6	0	0	1	+
7	1	1	0	-
8	0	0	0	-
9	0	1	0	+
10	1	1	1	+

3.1 a

Solution:

1. $P(A = 1|+) = \frac{3}{5}$
2. $P(B = 1|+) = \frac{2}{5}$
3. $P(C = 1|+) = \frac{4}{5}$
4. $P(A = 1|-) = \frac{2}{5}$
5. $P(B = 1|-) = \frac{2}{5}$
6. $P(C = 1|-) = \frac{1}{5}$

3.2 b

Use the estimate of conditional probabilities above to predict class label for a test sample $(A = 1, B = 1, C = 1)$:

Solution:

1. $P(Class = +) = \frac{3}{5}\frac{2}{5}\frac{4}{5} = \frac{24}{125}$
2. $P(Class = -) = \frac{2}{5}\frac{2}{5}\frac{1}{5} = \frac{4}{125}$

Because $P(Class = +) > P(Class = -)$, we assign this test point to class +

3.3 c

Compare $P(A = 1)$,$P(B = 1)$, and $P(A = 1, B = 1)$

Solution: $P(A = 1) = \frac{1}{2}.P(B = 1) = \frac{2}{5}.P(A = 1, B = 1) = \frac{1}{5}$. We see here that $P(A = 1) * P(B = 1) = \frac{1}{2} * \frac{2}{5} = \frac{2}{10} = \frac{1}{5} = P(A = 1, B = 1)$. A and B are independent because independence condition holds

3.4 d

Repeat the analysis in part (c) using $P(A=1)$,$P(B=0)$, and $P(A=1,B=0)$

Solution: $P(A=1)=\frac{1}{2}$. $P(B=0)=\frac{3}{5}$. $P(A=1,B=0)=\frac{3}{10}$. We test independence via $P(A=1)P(B=1)=\frac{1}{2}\frac{3}{5}=\frac{3}{10}=P(A=1,B=0)\rightarrow$ A and B are independent.

3.5 e

For independence, $P(A,B)=P(A)P(B)$ must hold. Hence to check that variables A and B and conditionally independent given the class, we check that $P(A=1,B=1|Class=+)=P(A=1|Class=+)P(B=1|Class=+)$

Solution: $P(A=1|Class=+)=\frac{3}{5}$. $P(B=1|Class=+)=\frac{2}{5}$. $P(A=1,B=1|Class=+)=\frac{1}{5}$. We test independence via $P(A=1|Class=+)P(B=1|Class=+)=\frac{3}{5}\frac{2}{5}=\frac{6}{25}\neq\frac{1}{5}=P(A=1,B=1|Class=+)\rightarrow$ A and B are conditional independent given the class.

4 Exercise 4: SVM

4.1 1

Solution: A larger margin means there is more distance between the closest point and the hyper-plane. With a larger margin, there is a decreased chance of over-fitting. In other words, you are in increasing the safety margin of making the wrong classification.

4.2 2

Show that, irrespective of the dimensionality of the data space, a data set consisting of just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyper-plane.
Solution: Let $x_1\in C^+$ and $x_2\in C^-$. Given the max-margin hyper-plane that classifies x_1 and x_2 is defined by

$$\min_{w,b}\frac{1}{2}||w||^2$$

s.t.

$$w^Tx_1+b-1=0$$

$$w^Tx_2+b+1=0$$

The Lagrangian is given by

$$L=\frac{1}{2}||w||^2+\alpha(w^Tx_1+b-1)+\beta(w^Tx_2+b+1)$$

$$\frac{\delta L}{\delta b} = 0 \to \alpha + \beta = 0 \to \alpha = -\beta$$
$$\frac{\delta L}{\delta w} = 0 \to w + \alpha x_1 + \beta x_2 = 0$$

Using the above result:
$$w = \beta(x_1 - x_2)$$

4.3 3

Primal Soft Margin SVM

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i$$

s.t

$$Y_i(w^T X_i + b) \geq 1 - \xi_i, i = 1, ..., m$$
$$\xi_i \geq 0, i = 1, ..., m$$

Solution: We first we rewrite the primal in standard form:

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i$$

s.t

$$(1 + \xi_i) - Y_i(w^T X_i + b) \leq 0, i = 1, ..., m$$
$$-\xi_i \leq 0, i = 1, ..., m$$

The Lagrangian here is:

$$L = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{m} \alpha_i[(1 - \xi_i) - Y_i(w^T X_i + b)] + \sum_{i=1}^{m} \beta_i[-\xi_i]$$

$$\frac{\delta L}{\delta \xi_i} = 0 \to C - \alpha_i - \beta_i = 0$$
$$\beta_i = C - \alpha_i$$
$$\frac{\delta L}{\delta b} = 0 \to \sum_{i=1}^{m} \alpha_i Y_i = 0$$
$$\frac{\delta L}{\delta w} = 0 \to 2\frac{1}{2}w - \sum_{i=1}^{m} \alpha_i Y_i X_i = 0$$
$$w = \sum_{i=1}^{m} \alpha_i Y_i X_i$$

Plugging these values back in to the Lagrangian, we get:

$$\frac{1}{2}||\sum_{i=1}^{m}\alpha_i Y_i X_i||^2+C\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{m}\alpha_i[(1-\xi_i)-Y_i(w^T X_i+\sum_{i=1}^{m}\alpha_i Y_i)]+\sum_{i=1}^{m}C-\alpha_i[-\xi_i]=$$

We simplify the first term using the information in class for the derivation of the hard-margin SVM dual:

$$\frac{1}{2}\sum_{i,j=1}^{m}Y_iY_j\alpha_i\alpha_j(X_i\cdot X_j)+C\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{m}\alpha_i[(1-\xi_i)-Y_i(w^T X_i+\sum_{i=1}^{m}\alpha_i Y_i)]+\sum_{i=1}^{m}C-\alpha_i[-\xi_i]=$$

Expanding and separating the summations where appropriate by linearity:

$$\frac{1}{2}\sum_{i,j=1}^{m}Y_iY_j\alpha_i\alpha_j(X_i\cdot X_j)+C\sum_{i=1}^{n}\xi_i+$$

$$\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\alpha_i\xi_i-\sum_{i=1}^{m}\alpha_i[Y_i(w^T X_i+\sum_{i=1}^{m}\alpha_i Y_i)]+$$

$$-\sum_{i=1}^{m}C\xi_i+\sum_{i=1}^{m}\xi_i\alpha_i=$$

Simplifying terms that cancel and combine:

$$\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\alpha_i\alpha_jY_iY_j(X_j\cdot X_j)$$

The dual becomes:

$$\min\max\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\alpha_i\alpha_jY_iY_j(X_j\cdot X_j)$$

s.t.

$$\sum_{i=1}^{m}\alpha_i y_i=0$$

$$o\leq\alpha_i\leq C$$