Machine Learning

July 20, 2016

Basic Concepts: Review

Example machine learning problem: Decide whether to play tennis at a given day.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Basic Concepts: Review

Example machine learning problem: Decide whether to play tennis at a given day.

Input Attributes
- or -
Input Variables
- or -
Features
- or -
Attributes

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Basic Concepts: Review

Example machine learning problem: Decide whether to play tennis at a given day.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Target Variable
- or -
Class Label
- or -
Goal
- or -
Output Variable

Basic Concepts: Review

Supervised Learning:

- Output variables (class labels) are given.
- The relationship between input and output is known.

Reinforced Learning:

- Output variables are not known, but actions are rewarded or punished.

Unsupervised Learning:

- Learn patterns from data without output variable or feedback.

(Semi-supervised Learning:)

- Only a small amount of data is labeled.

Basic Concepts: Review

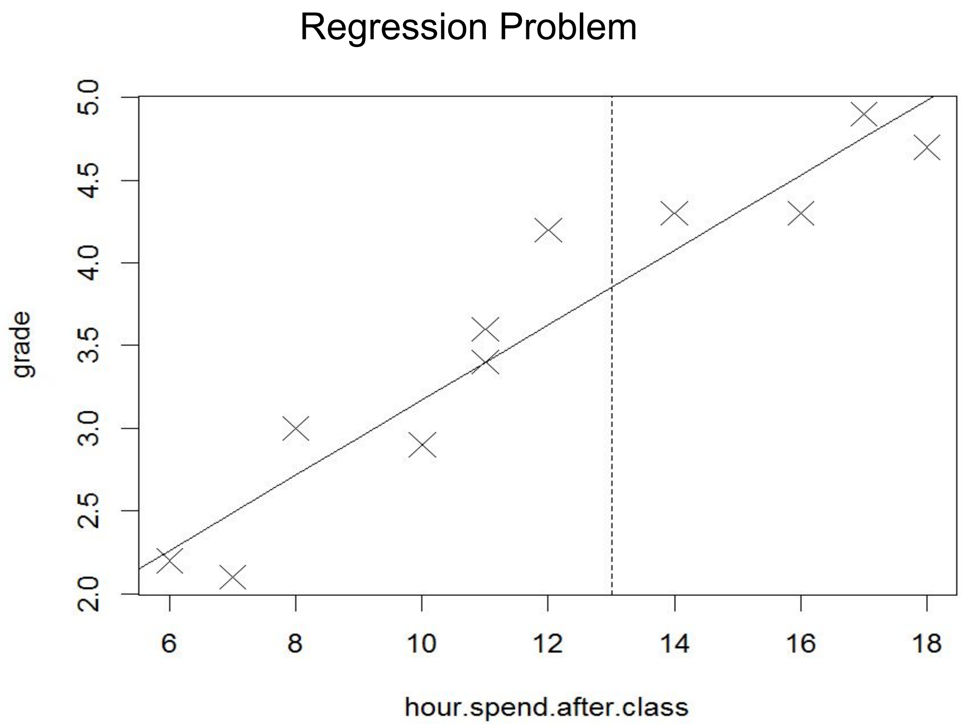
- **In Supervised Learning:**
    - Classification: Output variable takes a finite set of values (Categorical Variable).
    - Regression: Output variable is numeric (Continuous Variable).

- **In Unsupervised Learning:**
    - Clustering is a common approach.

Classification vs Regression

Hours spend After class	Grade
11	3.4
8	3.0
11	3.6
6	2.2
17	4.9
18	4.7
10	2.9
7	2.1
12	4.2
14	4.3
16	4.3

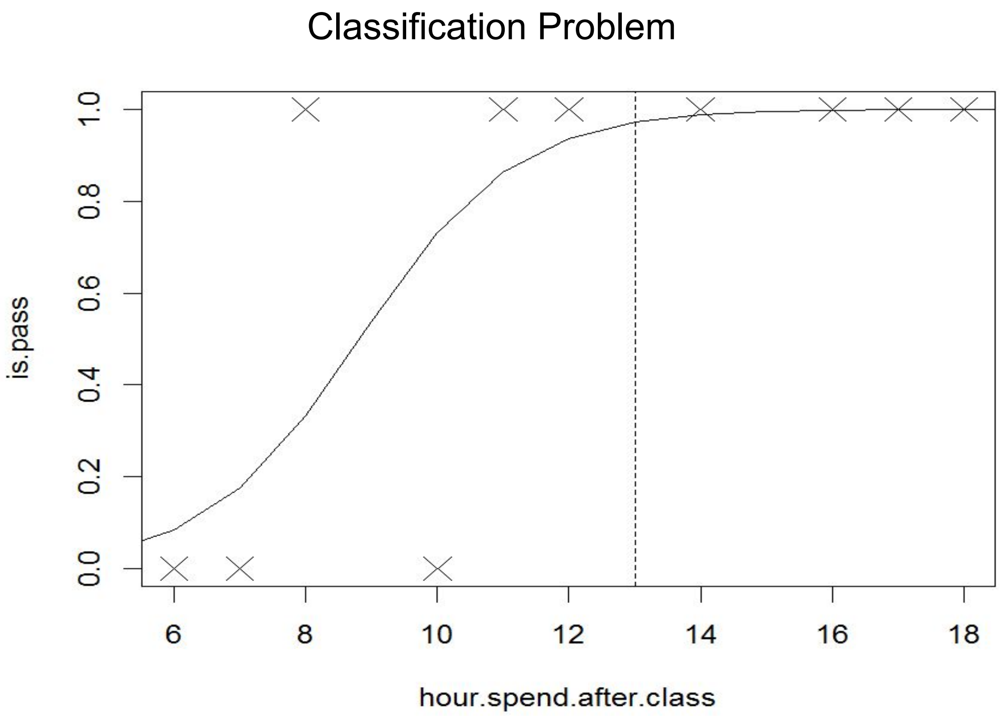
Given a new student S who spend 13 hours, what is the best guess of his/her grade?



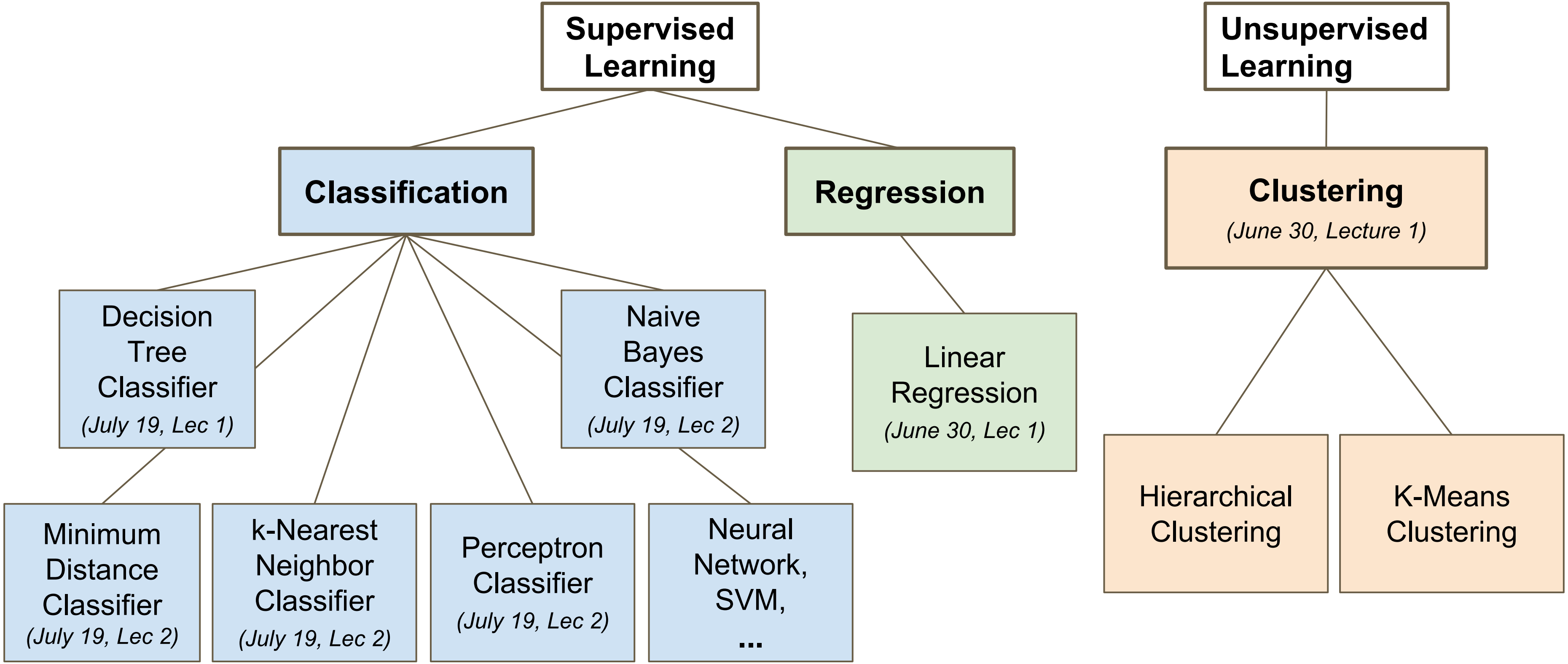
Classification vs Regression

Hours spend After class	Grade
11	Y
8	Y
11	Y
6	N
17	Y
18	Y
10	N
7	N
12	N
14	N
16	N

Given a new student S who spend 13 hours, how likely will he/she pass the class?



In CS171, we learned:

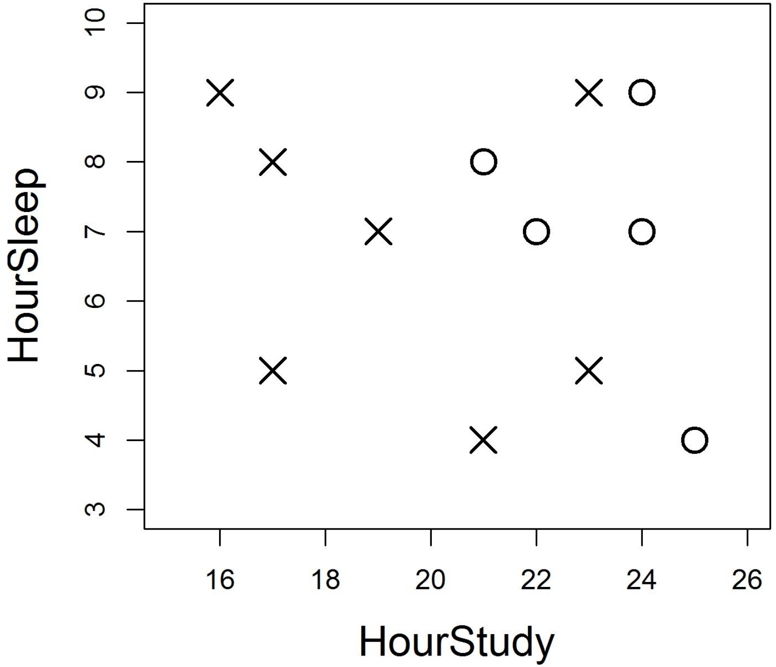


Note: Most classification methods can be applied to regression problems.

Decision Tree: Exercise 1

Consider the following set of training examples. There are two features: Number of hours a student spent studying (HourStudy), and the number of hours a student spent sleeping the night before the exam (HourSleep). The target variable is whether the student pass the class (Grade). The data is plotted on the right.

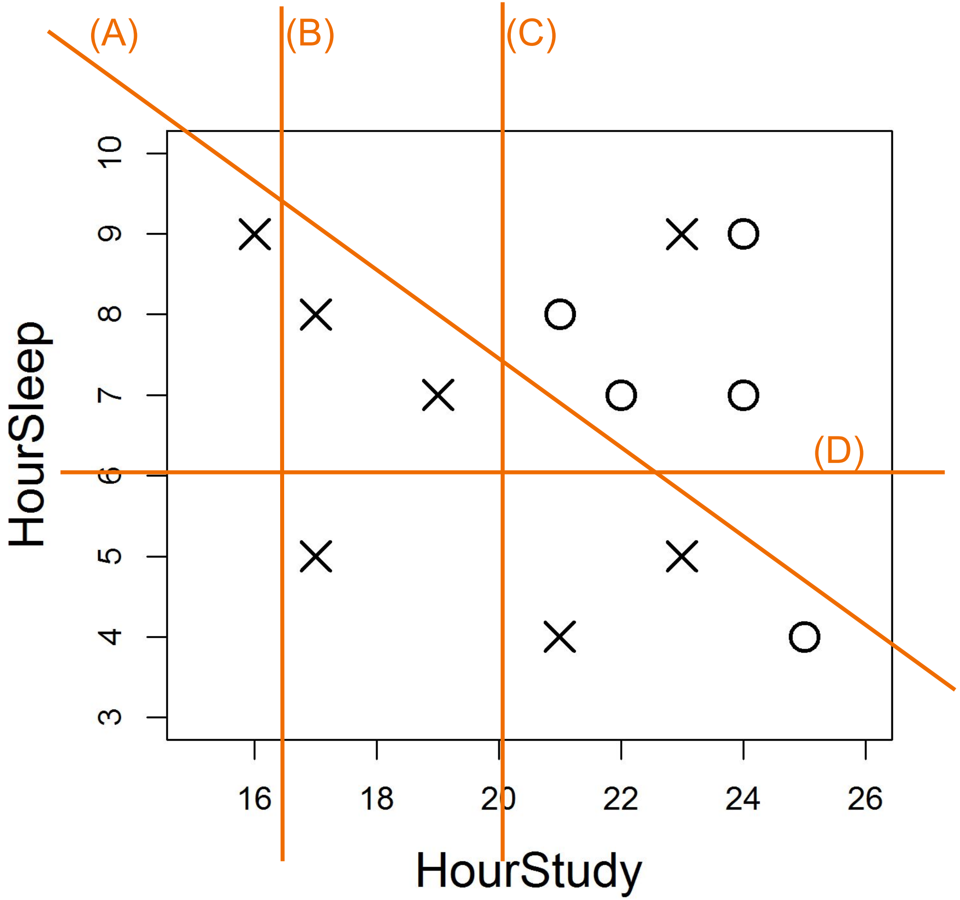
HourStudy	HourSleep	Grade
16	9	F
17	5	F
17	8	F
19	7	F
21	4	F
23	9	F
23	5	F
21	8	P
22	7	P
24	7	P
24	9	P
25	4	P



Decision Tree: Exercise 1

Use **Decision Tree Classifier**, which line best split the data as the first split?

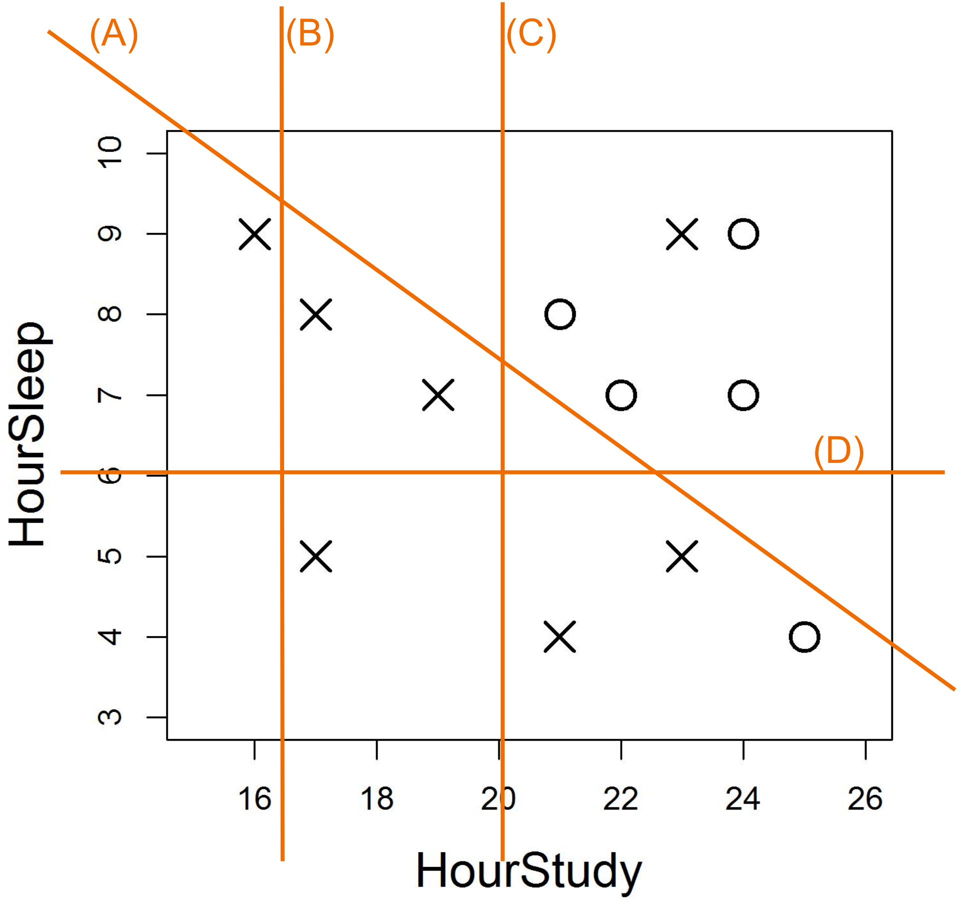
HourStudy	HourSleep	Grade
16	9	F
17	5	F
17	8	F
19	7	F
21	4	F
23	9	F
23	5	F
21	8	P
22	7	P
24	7	P
24	9	P
25	4	P



Decision Tree: Exercise 1

Use **Decision Tree Classifier**, which line best split the data as the first split?

HourStudy	HourSleep	Grade
16	9	F
17	5	F
17	8	F
19	7	F
21	4	F
23	9	F
23	5	F
21	8	P
22	7	P
24	7	P
24	9	P
25	4	P



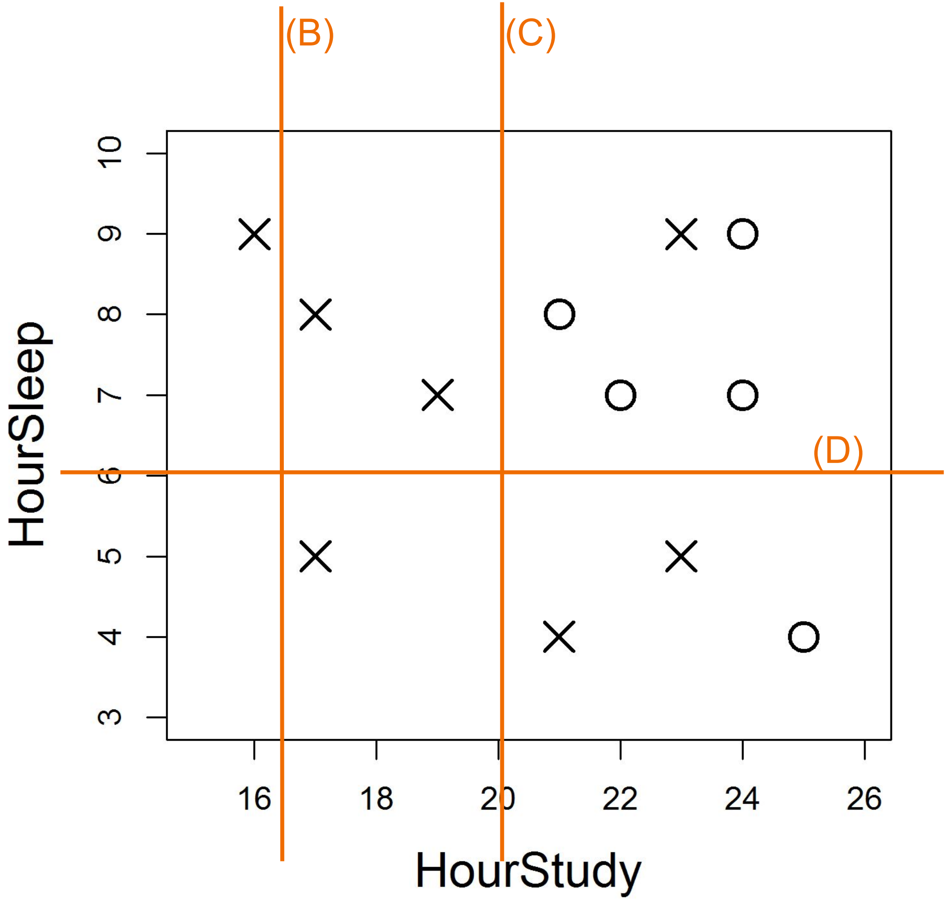
(A) is incorrect because when using decision tree we are splitting 1 variable at a time. Decision boundaries have to be perpendicular to x or y axis.

Decision Tree: Exercise 1

Use **Decision Tree Classifier**, which line best split the data as the first split?

HourStudy	HourSleep	Grade
16	9	F
17	5	F
17	8	F
19	7	F
21	4	F
23	9	F
23	5	F
21	8	P
22	7	P
24	7	P
24	9	P
25	4	P

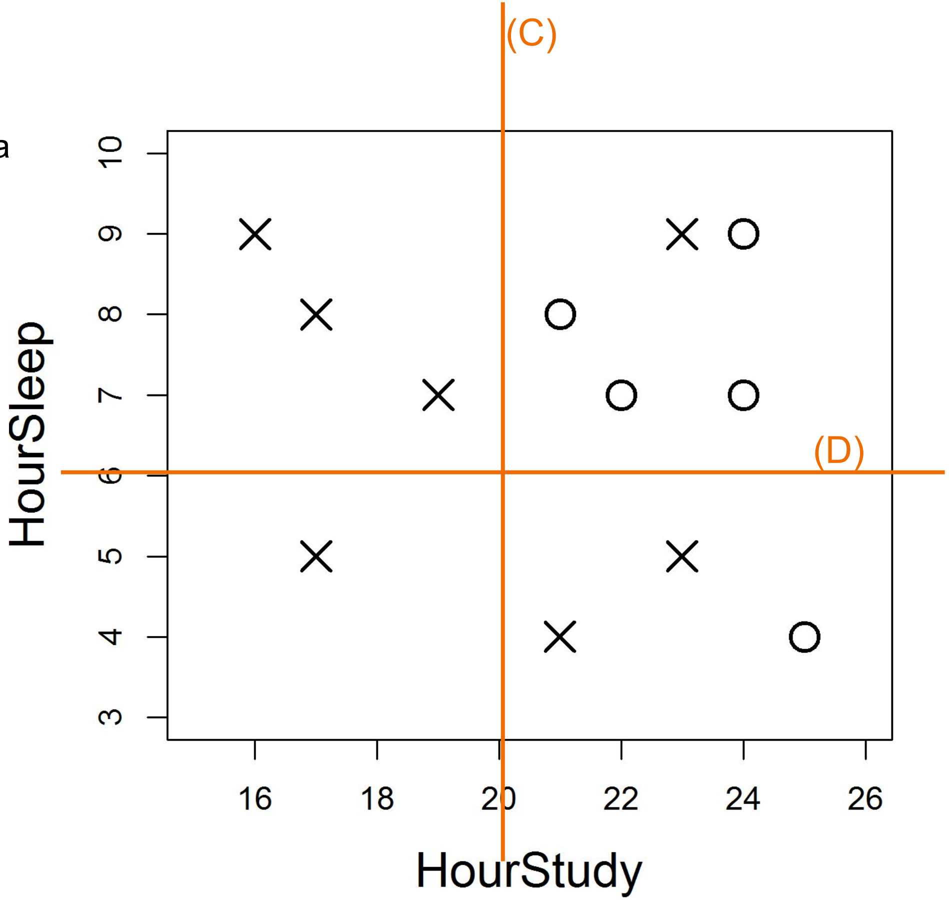
(B) is not a good split because it clearly doesn't differentiate the dataset.



Decision Tree: Exercise 1

Use **Decision Tree Classifier**, which line best split the da as the first split?

HourStudy	HourSleep	Grade
16	9	F
17	5	F
17	8	F
19	7	F
21	4	F
23	9	F
23	5	F
21	8	P
22	7	P
24	7	P
24	9	P
25	4	P



(C) and (D) **can** both be reasonable splits. We have to examine their **entropy** values after the split.

In binary case, $H(p) = -p \log p - (1-p) \log (1-p)$

Smaller entropy after splits $\Longrightarrow$ Greater information gain

Decision Tree: Exercise 1

In binary case, H(p) = - p log p - (1-p) log (1-p)

$$H(C_{\text{left}}) = 0$$

Pick (C)

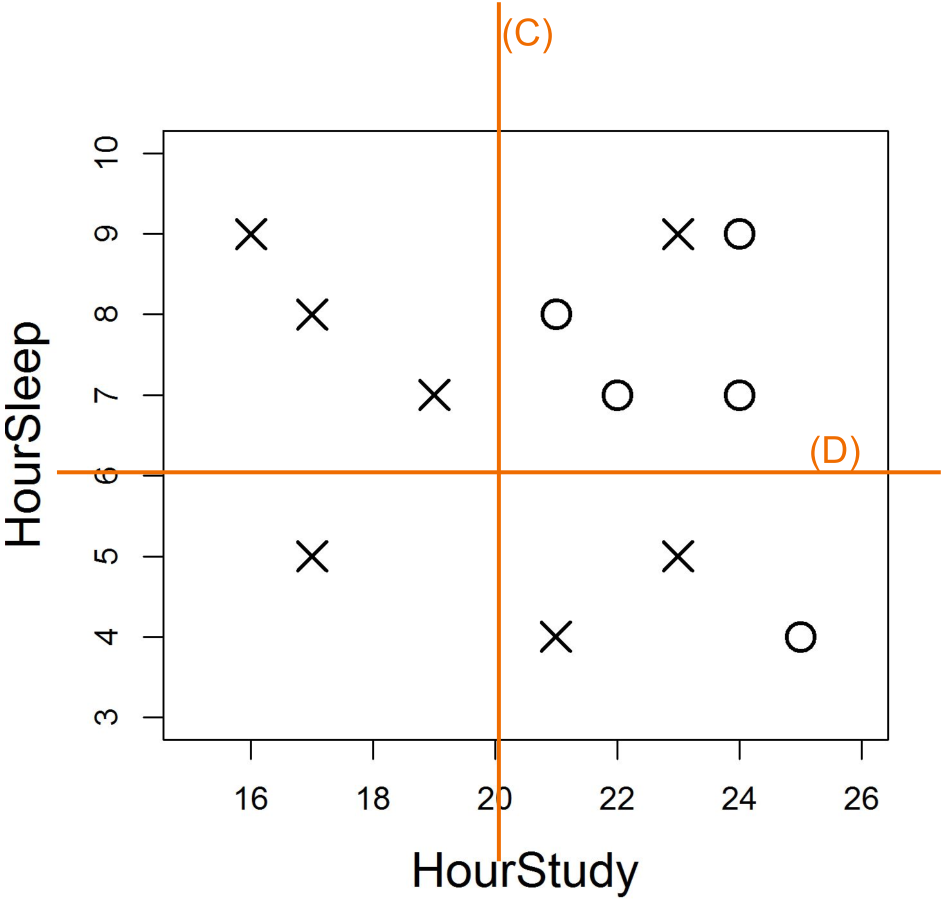
$$H(C_{\text{right}}) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = 0.95$$

$$H(C) = \frac{4}{12}H(C_{\text{left}}) + \frac{8}{12}H(C_{\text{right}}) = \underline{0.63}$$

$$H(D_{\text{top}}) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1$$

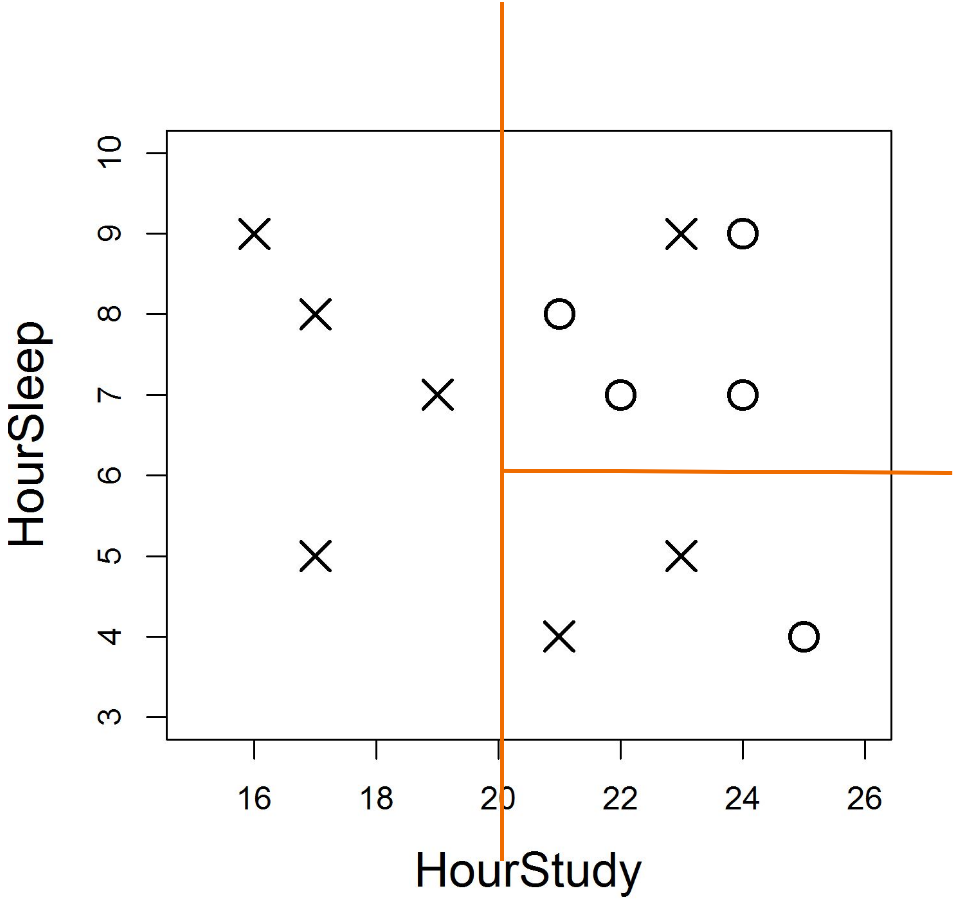
$$H(D_{\text{bottom}}) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.81$$

$$H(D) = \frac{8}{12}H(D_{\text{top}}) + \frac{4}{12}H(D_{\text{bottom}}) = \underline{0.94}$$



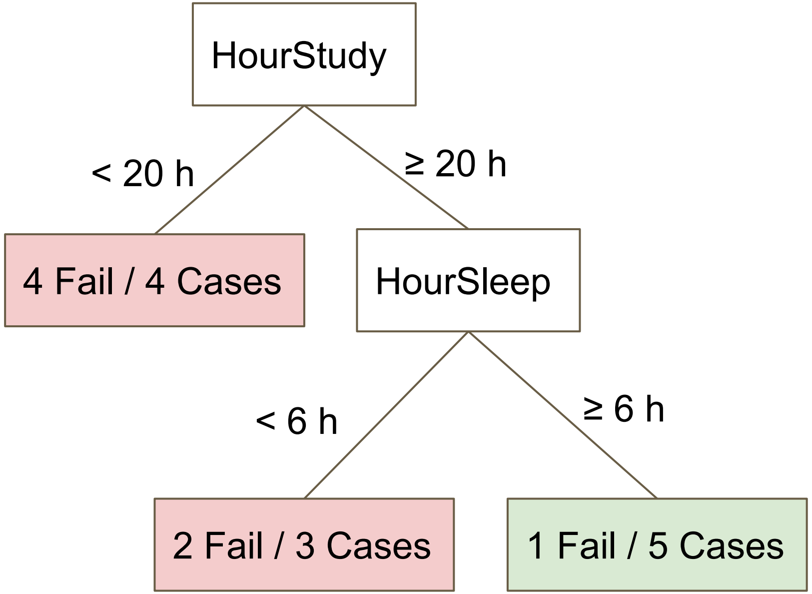
Decision Tree: Exercise 2

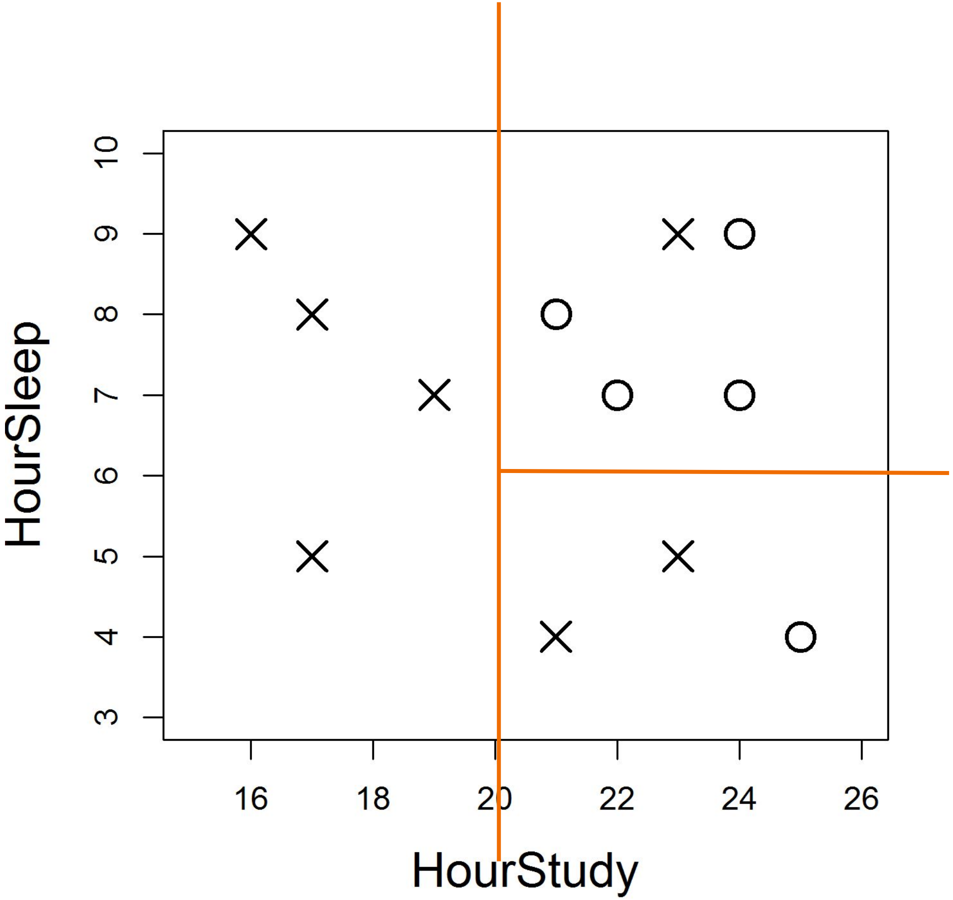
Assume the first and second splits are shown on the figure. Rewrite the splits using the tree representation:



Decision Tree: Exercise 2

Assume the first and second splits are shown on the figure. Rewrite the splits using the tree representation:

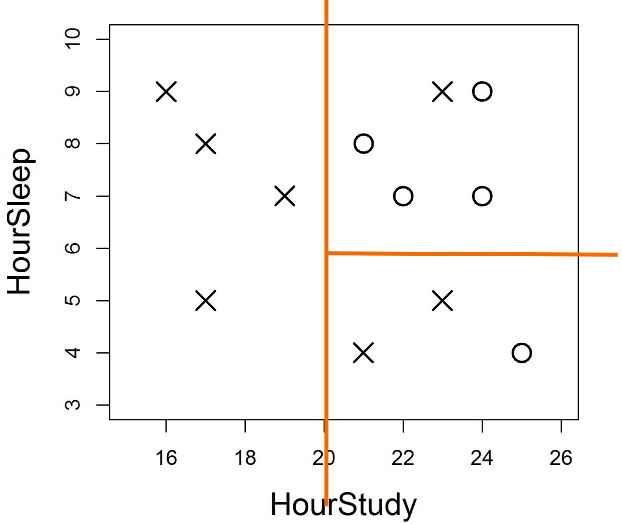


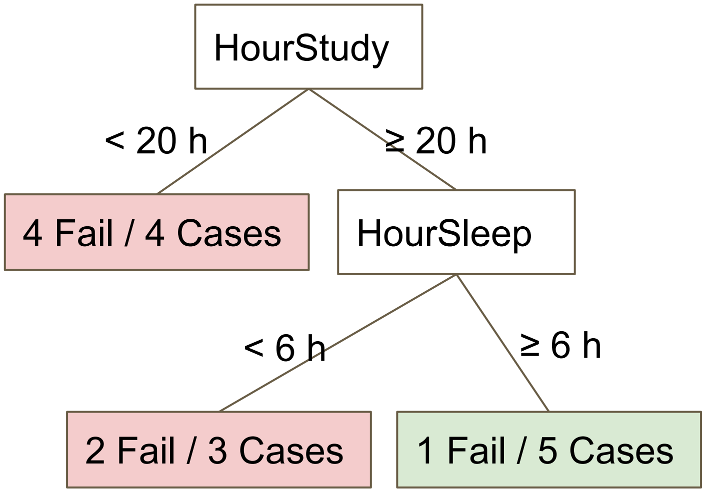


Decision Tree: Exercise 3

Classify the following test data cases. You should be able to obtain the predictions using either representation.

Student	HourStudy	HourSleep	Pass?
Alice	16	9	
Bob	26	5	
Charlie	21	8	

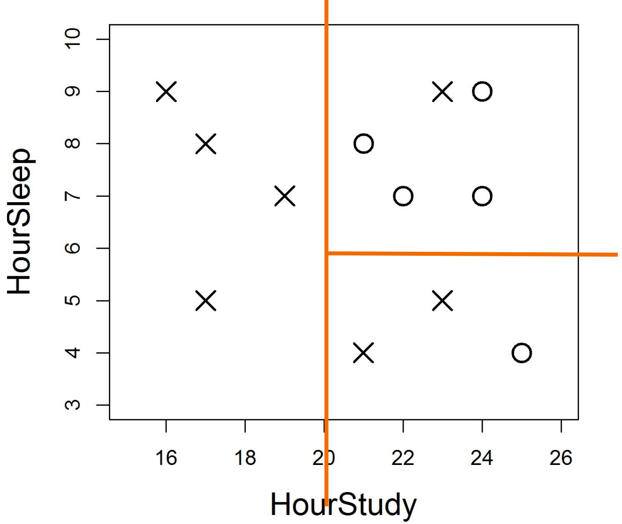


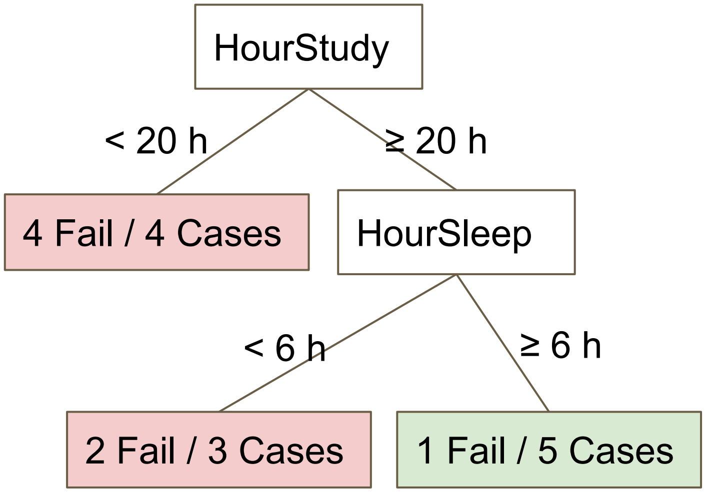


Decision Tree: Exercise 3

Classify the following test data cases. You should be able to obtain the predictions using either representation.

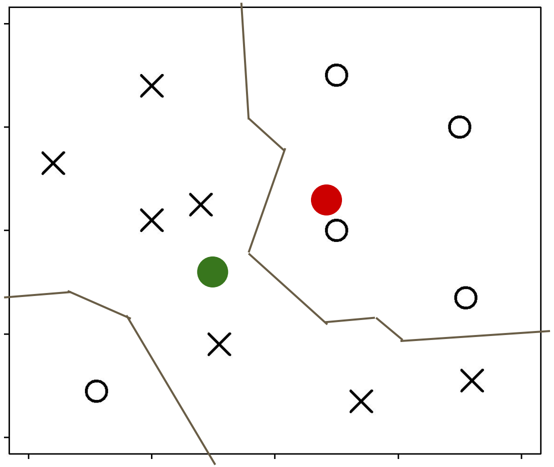
Student	HourStudy	HourSleep	Pass?
Alice	16	9	F
Bob	26	5	F
Charlie	21	8	T



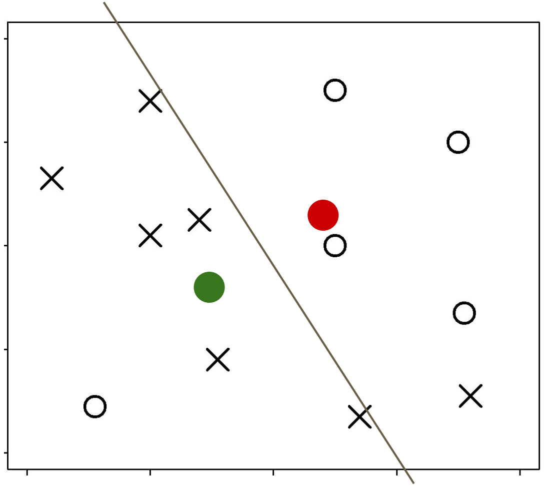


Decision Boundary: Exercise

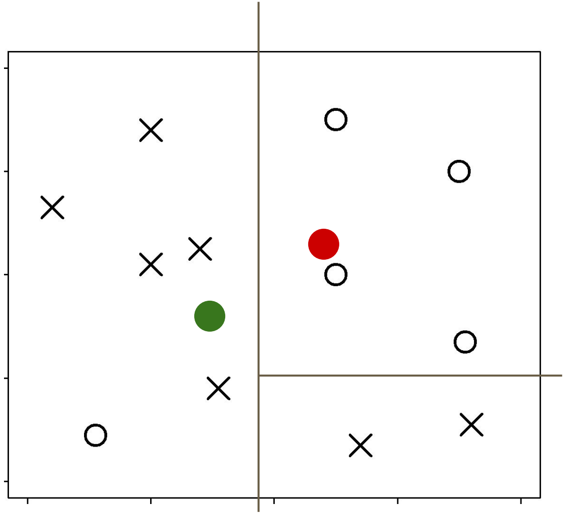
Match the decision boundary with the most probable classifiers. (Mean for each class is shown as red/green circle)



(A)



(B)

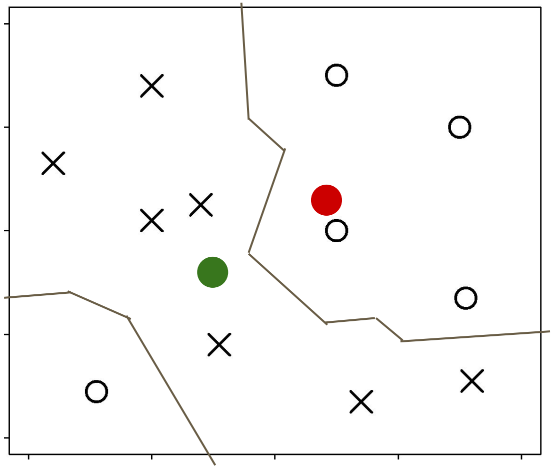


(C)

(1) Decision Tree Classifier; (2) Minimum Distance Classifier; (3) Nearest Neighbor Classifier.

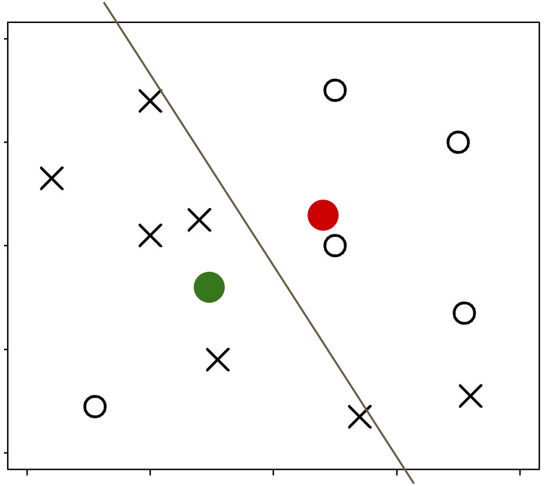
Decision Boundary: Exercise

Match the decision boundary with the most probable classifiers. (Mean for each class is shown as red/green circle)



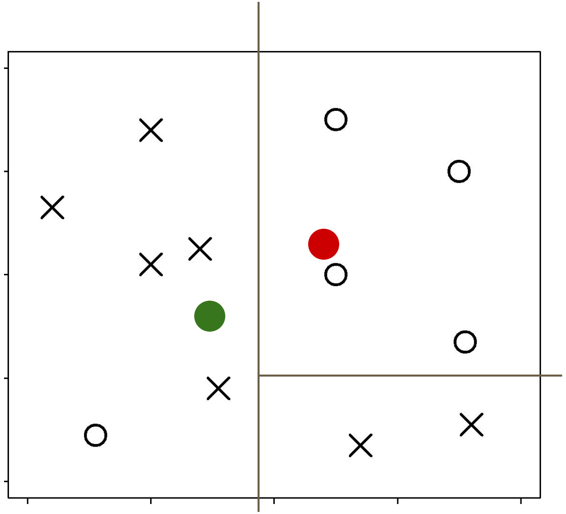
(A)

(3) Nearest Neighbor Classifier



(B)

(2) Minimum Distance Classifier



(C)

(1) Decision Tree Classifier

Naive Bayes Classifier: Example

Consider the following set of training examples. A and B are features and Y is the target variable. Each row indicates the values observed, and how many times that set of values was observed. For example, (t, t, 1) was observed 3 times, while (t, t, 0) was never observed.

A	B	Y	Count
t	t	1	3
t	f	1	2
f	t	1	1
f	f	1	2
t	t	0	0
t	f	0	1
f	t	0	1
f	f	0	2

In general:

$$P(Y|X_1, X_2, \dots) = \frac{P(X_1, X_2, \dots, Y)}{P(X_1, X_2, \dots)} = \frac{P(X_1, X_2, \dots|Y)P(Y)}{P(X_1, X_2, \dots)} \qquad \text{(Bayes' rule)}$$

$$= \alpha \ P(X_1, X_2, \dots|Y)P(Y) \qquad (\text{let } \alpha = \frac{1}{P(X_1, X_2, \dots)}).$$

$$= \alpha \ P(Y)P(X_1|Y)P(X_2|P)\dots \qquad \text{(Features are independent given the class)}$$

Apply to this problem:

$$P(Y|A, B) = \alpha \ P(A|Y)P(B|Y)P(Y)$$

We just need to calculate P(A|Y) P(B|Y) and P(Y)

This is a variation of problem 1 in http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/homework/hw5-sol.pdf

Naive Bayes Classifier: Example

Consider the following set of training examples. A and B are features and Y is the target variable. Each row indicates the values observed, and how many times that set of values was observed. For example, (t, t, 1) was observed 3 times, while (t, t, 0) was never observed.

A	B	Y	Count
t	t	1	3
t	f	1	2
f	t	1	1
f	f	1	2
t	t	0	0
t	f	0	1
f	t	0	1
f	f	0	2

$$P(Y|A,B) = \alpha\ P(A|Y)P(B|Y)P(Y)$$

Eg.
P(A = f | Y = 1) = α 3/8 ; P(B = t | Y = 1) = α 4/8
P(Y = 1) = α 8/12

Given a test data case (f, t, ?), what is the most probable Y value?

This is a variation of problem 1 in http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/homework/hw5-sol.pdf

Naive Bayes Classifier: Example

Consider the following set of training examples. A and B are features and Y is the target variable. Each row indicates the values observed, and how many times that set of values was observed. For example, (t, t, 1) was observed 3 times, while (t, t, 0) was never observed.

A	B	Y	Count
t	t	1	3
t	f	1	2
f	t	1	1
f	f	1	2
t	t	0	0
t	f	0	1
f	t	0	1
f	f	0	2

$$P(Y|A,B) = \alpha\ P(A|Y)P(B|Y)P(Y)$$

Eg.
P(A = f | Y = 1) = α 3/8 ; P(B = t | Y = 1) = α 4/8
P(Y = 1) = α 8/12

Given a test data case (f, t, ?), what is the most probable Y value?

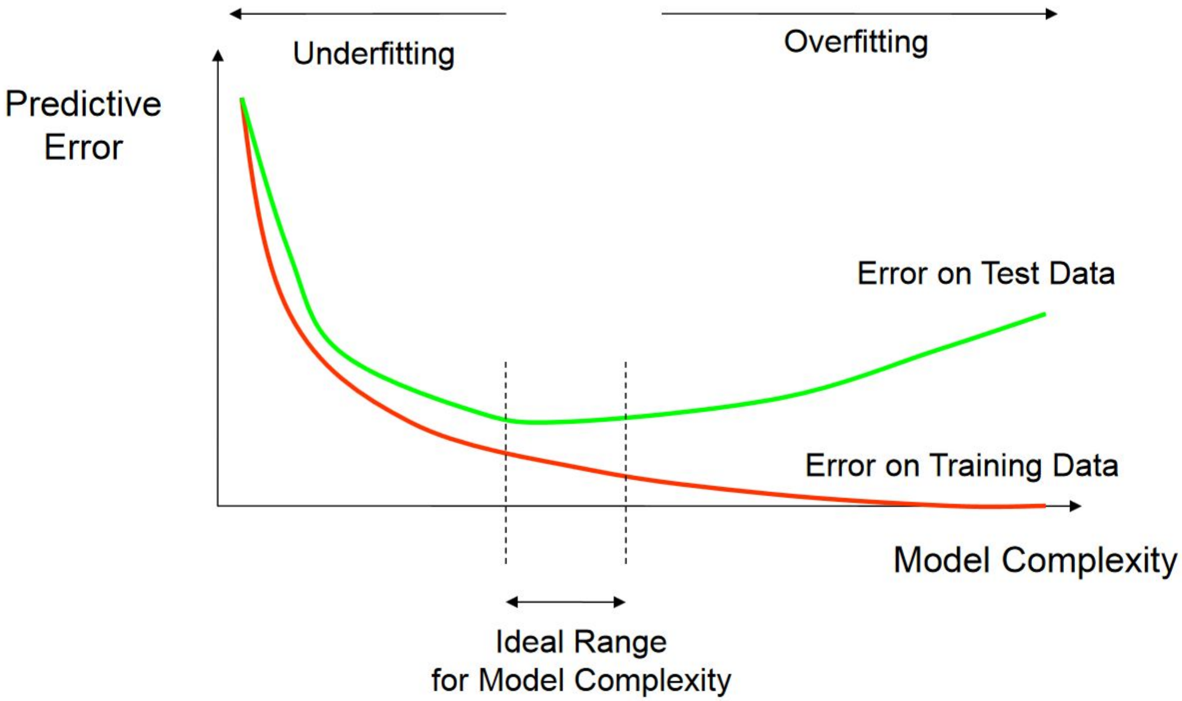
P(Y = 1|A = f, B = t) = α P(A = f | Y = 1) P(B = t | Y = 1) P(Y = 1) = α 3/8*4/8*8/12
= α 1/8
P(Y = 0|A = f, B = t) = α P(A = f | Y = 0) P(B = t | Y = 0) P(Y = 0) = α 3/4*1/4*4/12
= α 1/16

P(Y = 1|A = f, B = t) > P(Y = 0|A = f, B = t); The prediction is Y = 1.

This is a variation of problem 1 in http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/homework/hw5-sol.pdf

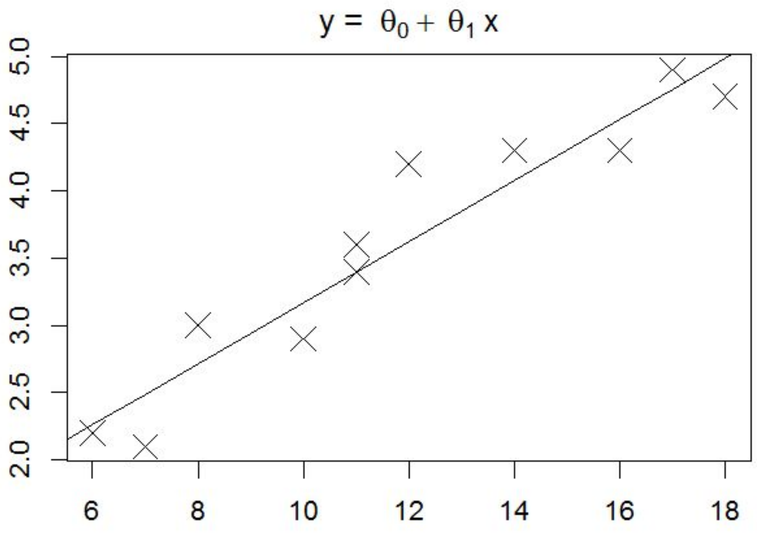
Bias vs. Variance (Underfitting vs. Overfitting): Review

Underfitting: Error is caused by model bias.
Overfitting: Error is caused by data variance. *(Slide 45-55, Lec1, July 19).*

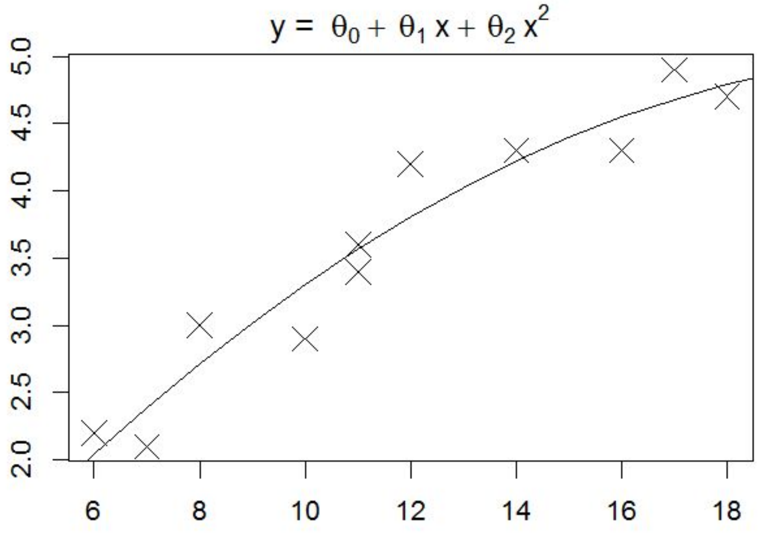


Bias vs. Variance (Underfitting vs. Overfitting): Review

Model complexity in linear regression can be characterized by the number of parameters in the polynomial.

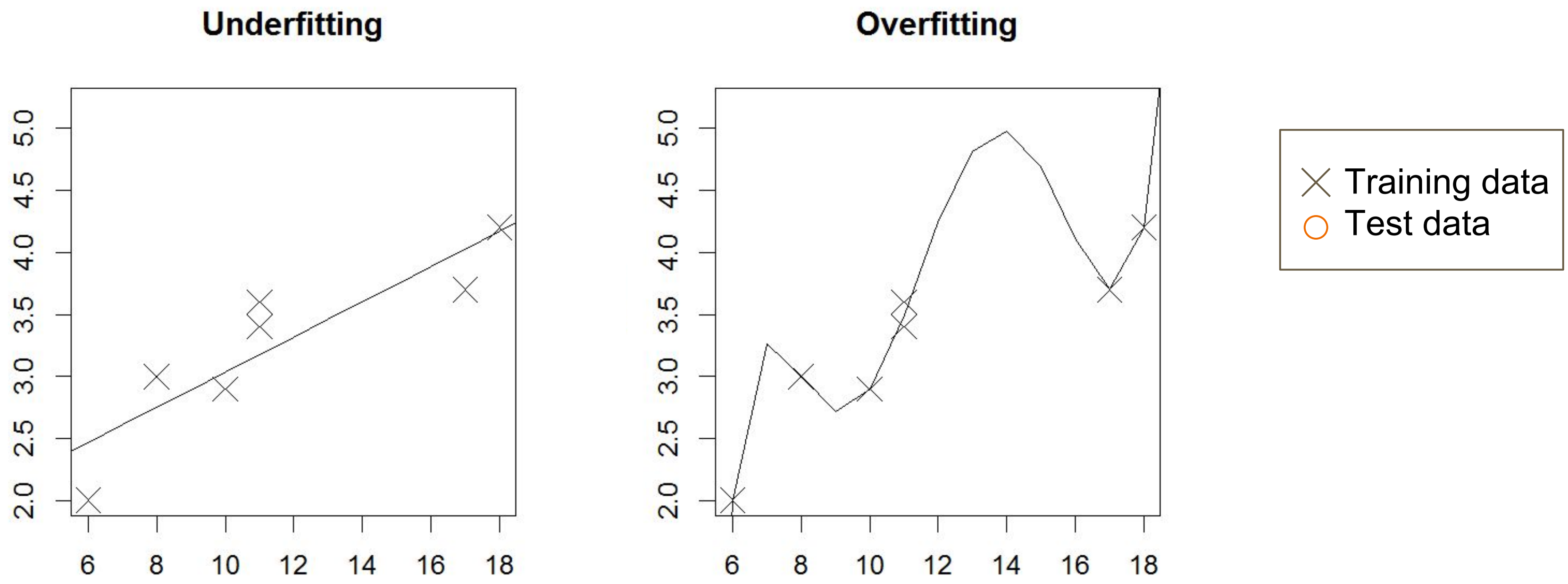


MSE = 0.0806



MSE = 0.0602

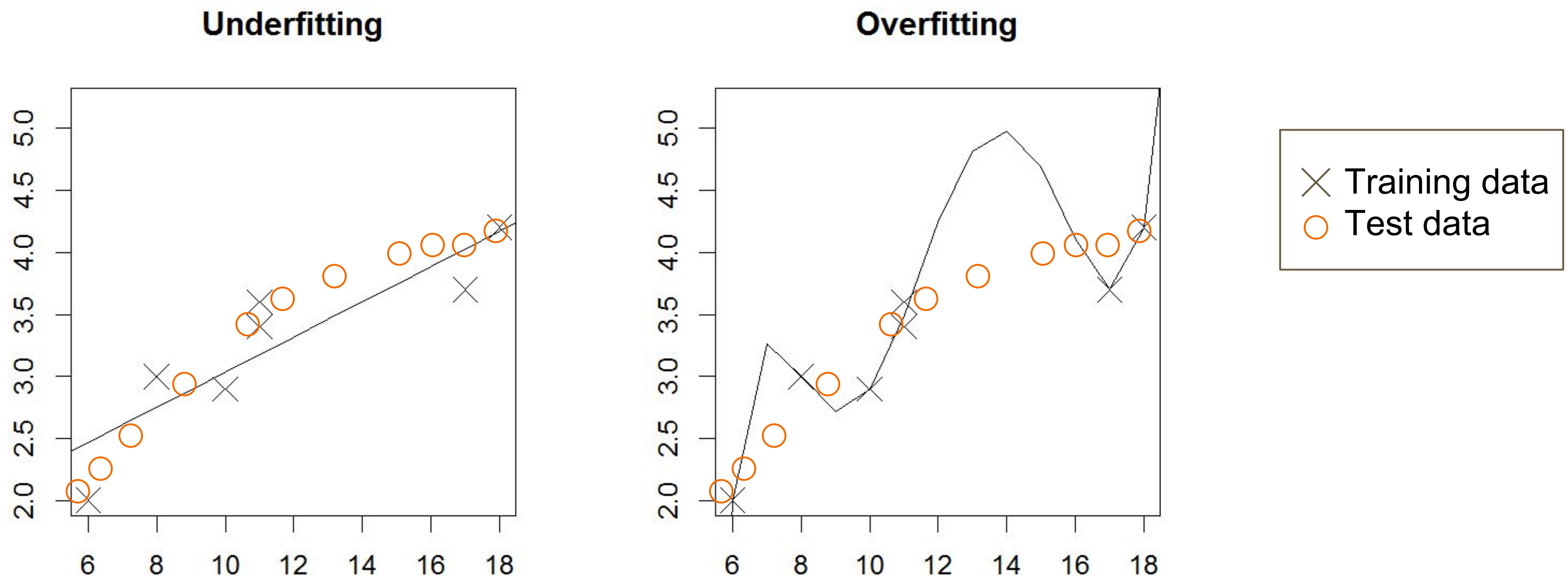
Bias vs. Variance (Underfitting vs. Overfitting): Example



On the left: Linear regression (2 parameters). On the right: Polynomial regression (6 parameters).

Polynomial regression with 6 parameters is more complex than linear regression with 2 parameters, thus achieves smaller training error. (Assume the error measure is MSE = mean squared distance to the fitted line)

Bias vs. Variance (Underfitting vs. Overfitting): Example

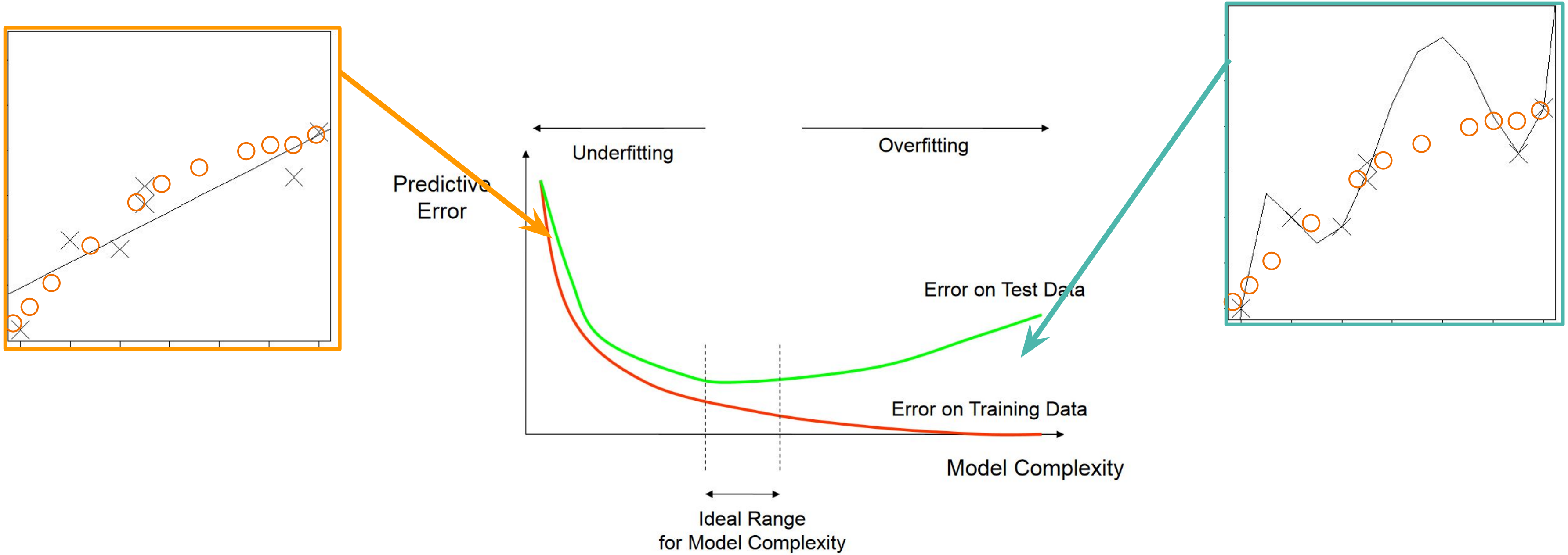


However, when we used the fitted line to predict the values of the test data, polynomial model with 6 parameters suffers. It is because the model overfits the training data. Linear model suffers too (to a lesser extent) because it is too simple for the data.

Bias vs. Variance (Underfitting vs. Overfitting): Review

Underfitting: Error is caused by model bias.

Overfitting: Error is caused by data variance. *(Slide 45-55, Lec1, July 19).*



Nearest Neighbor Classifier & Cross Validation

Consider this training data set with 9 students' final scores and class grade. The single feature is **Final Score**, and class labels (**Grade**) are A, B, or C. *(This is a variation of Question 1, Final Exam, Fall 2014).*

Student	1	2	3	4	5	6	7	8	9
Final Score	53	59	70	79	84	87	91	93	99
Grade	B	C	B	B	A	B	A	A	A

Using 1-Nearest Neighbor, what class label would be assigned to a new student, who has Final Score = 86?

Using 3-Nearest Neighbor, what class label would be assigned to a new student, who has Final Score = 86?

Nearest Neighbor Classifier & Cross Validation

Consider this training data set with 9 students' final scores and class grade. The single feature is **Final Score**, and class labels (**Grade**) are A, B, or C. *(This is a variation of Question 1, Final Exam, Fall 2014).*

Student	1	2	3	4	5	6	7	8	9
Final Score	53	59	70	79	84	87	91	93	99
Grade	B	C	B	B	A	B	A	A	A

Using 1-Nearest Neighbor, what class label would be assigned to a new student, who has Final Score = 86? **B**

Using 3-Nearest Neighbor, what class label would be assigned to a new student, who has Final Score = 86? **A**

Nearest Neighbor Classifier & Cross Validation

Consider this training data set with 9 students' final scores and class grade. The single feature is **Final Score**, and class labels (**Grade**) are A, B, or C. *(This is a variation of Question 1, Final Exam, Fall 2014).*

Student	1	2	3	4	5	6	7	8	9
Final Score	53	59	70	79	84	87	91	93	99
Grade	B	C	B	B	A	B	A	A	A

Using 1-Nearest Neighbor and 3-fold Cross-Validation, what is the cross-validated accuracy of 1-Nearest Neighbor on this training set? (The validation partitions are given to you as Partition 1 = {1,4,7}; Partition 2 = {2,5,8}; Partition 3 = {3,6,9})

Nearest Neighbor Classifier & Cross Validation

Consider this training data set with 9 students' final scores and class grade. The single feature is **Final Score**, and class labels (**Grade**) are A, B, or C. *(This is a variation of Question 1, Final Exam, Fall 2014).*

Student	1	2	3	4	5	6	7	8	9
Final Score	53	59	70	79	84	87	91	93	99
Grade	B	C	B	B	A	B	A	A	A

Color Labels

"Training Set"

Validation Set

Using 1-Nearest Neighbor and 3-fold Cross-Validation, what is the cross-validated accuracy of 1-Nearest Neighbor on this training set? (The validation partitions are given to you as Partition 1 = {1,4,7}; Partition 2 = {2,5,8}; Partition 3 = {3,6,9})

Partition 1: Student 1's nearest neighbor is Student 2, predict C. - **Incorrect.**
Student 4's nearest neighbor is Student 5, predict A. - **Incorrect.**
Student 7's nearest neighbor is Student 8, predict A. - **Correct.**

Accuracy of Partition 1 = 1/3

Nearest Neighbor Classifier & Cross Validation

Consider this training data set with 9 students' final scores and class grade. The single feature is **Final Score**, and class labels (**Grade**) are A, B, or C. *(This is a variation of Question 1, Final Exam, Fall 2014).*

Student	1	2	3	4	5	6	7	8	9
Final Score	53	59	70	79	84	87	91	93	99
Grade	B	C	B	B	A	B	A	A	A

Color Labels
"Training Set"
Validation Set

Using 1-Nearest Neighbor and 3-fold Cross-Validation, what is the cross-validated accuracy of 1-Nearest Neighbor on this training set? (The validation partitions are given to you as Partition 1 = {1,4,7}; Partition 2 = {2,5,8}; Partition 3 = {3,6,9})

Partition 2: Student 2's nearest neighbor is Student 1, predict B. - **Incorrect.**
Student 5's nearest neighbor is Student 6, predict B. - **Incorrect.**
Student 8's nearest neighbor is Student 7, predict A. - **Correct.**

Accuracy of Partition 1 = 1/3
Accuracy of Partition 2 = 1/3

Nearest Neighbor Classifier & Cross Validation

Consider this training data set with 9 students' final scores and class grade. The single feature is **Final Score**, and class labels (**Grade**) are A, B, or C. *(This is a variation of Question 1, Final Exam, Fall 2014).*

Student	1	2	3	4	5	6	7	8	9
Final Score	53	59	70	79	84	87	91	93	99
Grade	B	C	B	B	A	B	A	A	A

Color Labels

"Training Set"

Validation Set

Using 1-Nearest Neighbor and 3-fold Cross-Validation, what is the cross-validated accuracy of 1-Nearest Neighbor on this training set? (The validation partitions are given to you as Partition 1 = {1,4,7}; Partition 2 = {2,5,8}; Partition 3 = {3,6,9})

Partition 3: Student 3's nearest neighbor is Student 4, predict B. - **Correct.**
Student 6's nearest neighbor is Student 5, predict A. - **Incorrect.**
Student 9's nearest neighbor is Student 8, predict A. - **Correct.**

Accuracy of Partition 1 = 1/3
Accuracy of Partition 2 = 1/3
Accuracy of Partition 3 = 2/3

Cross-validated Accuracy =
1/3 * (1/3+1/3+2/3) = 4/9

Nearest Neighbor Classifier & Cross Validation : Verify at home

Consider this training data set with 9 students' final scores and class grade. The single feature is **Final Score**, and class labels (**Grade**) are A, B, or C. *(This is a variation of Question 1, Final Exam, Fall 2014).*

Student	1	2	3	4	5	6	7	8	9
Final Score	58	59	70	79	84	87	91	93	99
Grade	B	C	B	B	A	B	A	A	A

Using **3-Nearest Neighbor** and 3-fold Cross-Validation, what is the cross-validated accuracy of 3-Nearest Neighbor on this training set? (The validation partitions are given to you as Partition 1 = {1,4,7}; Partition 2 = {2,5,8}; Partition 3 = {3,6,9})

Accuracy of Partition 1 (1,4,7 as validation set) = 1
Accuracy of Partition 2 (2,5,8 as validation set) = 1/3
Accuracy of Partition 3 (3,6,9 as validation set) = 2/3

Cross-validated Accuracy =1/3 * (1+1/3+2/3) = 2/3