Introduction to Machine Learning

Tadao Hoshino （星野匡郎）

Econometrics II: ver. 2019 Spring Semester

Types of Machine Learning Algorithms

Types of Machine Learning Algorithms

- Machine Learning algorithms are classified into three categories:
  1. Supervised Learning
  2. Unsupervised Learning
  3. Reinforcement Learning (out of the scope of this course)

Supervised Learning

- Two main areas where supervised learning is used: *classification* and *regression*.
- Usually, in supervised learning, we have two datasets, training data and test data:
    - Training data (teacher data): $(\mathbf{X}^{\text{train}}, \Upsilon^{\text{train}})$,
    - Test data: $(\mathbf{X}^{\text{test}}, \Upsilon^{\text{test}})$,

  where $\mathbf{X}$ is a vector of input variables (explanatory variables), and Υ is a response variable (dependent variable).
- In real-world applications, the value of Υ^{test} are unknown, while $\mathbf{X}^{\text{test}}$ is available.

Supervised Learning (cont.)

- Our task is to build a function $f(\mathbf{X})$ that generates the predicted value of Y using the training data $(\mathbf{X}^{\text{train}}, Y^{\text{train}})$:

$$\widehat{Y} = f(\mathbf{X}),$$

  and predict the value of Y^{test} by $\widehat{Y}^{\text{test}} = f(\mathbf{X}^{\text{test}})$ as accurately as possible.
- When Y is continuous = Regression problem
- When Y is categorical = (Supervised) classification problem[1]

[1] For classification problems, Y is particularly referred to as the "label".

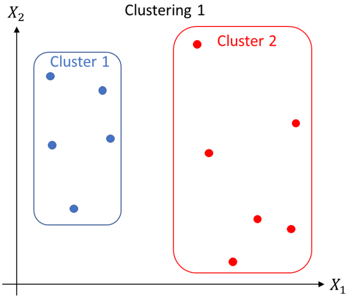
Unsupervised Learning

- Unsupervised learning is a type of machine learning technique based on datasets consisting only of input variables without responses.

  $\Rightarrow$ Unsupervised leaning does not have any known answers.
- Mostly, unsupervised learning focuses on two main areas: *clustering* and *dimension reduction*.
  - Clustering: classify a set of "unlabeled" objects into groups (clusters) based on some similarity measures.
  - Dimension reduction: summarize the information contained in large datasets ("Big Data") into a few synthetic variables; e.g., principal component analysis, variable selection, etc.[2]

[2] We will discuss more about dimension reduction techniques in the next lecture.

Clustering analysis.



* There are potentially multiple clustering results from the same dataset.

Reinforcement Learning

- = Dynamic Programming: which action should be taken in the current state to maximize future reward.
- A well-designed reinforcement learning algorithm sometimes overwhelms human experts.

Evaluation of Supervised Learning Algorithms

Accuracy and Error

- Consider a binary classification problem:
  - $\mathbf{X}$: input variables
  - $Y \in \{0, 1\}$: response variable
- Using a training dataset $\{(\mathbf{X}_i^{\text{train}}, Y_i^{\text{train}}) : 1 \le i \le N\}$, we compute a classification function $s(\mathbf{X})$ such that if $s(\mathbf{X}) \ge c$, we predict the response as $\widehat{Y} = 1$ (otherwise, $\widehat{Y} = 0$), i.e.,

$$\widehat{Y} = \mathbf{1}\{s(\mathbf{X}) \ge c\},$$

  where c is a pre-specified cut-off value.
- The function $s(\mathbf{X})$ can be obtained, for example, by a linear regression or a logistic regression, as described later.
- For a given test dataset $\{(\mathbf{X}_i^{\text{test}}, Y_i^{\text{test}}) : 1 \le i \le n\}$, we would like to predict the value of Y_i^{test} by $\widehat{Y}_i^{\text{test}}$ as accurately as possible.

Accuracy

- Note that for each test data point $\mathbf{X}_i^{\text{test}}$, the result of the prediction necessarily falls into one of the following four cases:

	True state = 1	**True state = 0**
Predicted state = 1	True Positive	False Positive
Predicted state = 0	False Negative	True Negative

- A natural measure for evaluating the performance of the classifier is the accuracy (i.e., hit rate):

$$\begin{aligned}\text{Accuracy} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i^{\text{test}} = \widehat{Y}_i^{\text{test}}\} \\ &= \frac{\#\text{TP} + \#\text{TN}}{\#\text{TP} + \#\text{TN} + \#\text{FP} + \#\text{FN}},\end{aligned}$$

where $\#\text{TP} = \sum_{i=1}^{n} \mathbf{1}\{Y_i^{\text{test}} = 1, \widehat{Y}_i^{\text{test}} = 1\}$, and the other terms are defined similarly.

Accuracy

- Note that the accuracy measure depends on the cut-off c, which needs to be determined beforehand.
- One might think that c should be set to a value that maximizes the accuracy.
- However, a drawback of this approach is that how the classifier is sensitive/robust to c is obscured.

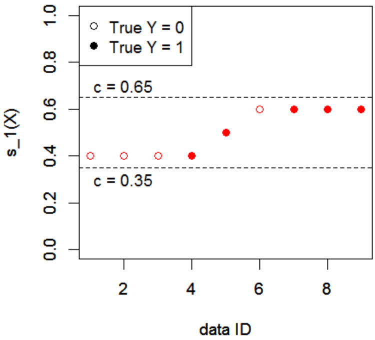
Accuracy

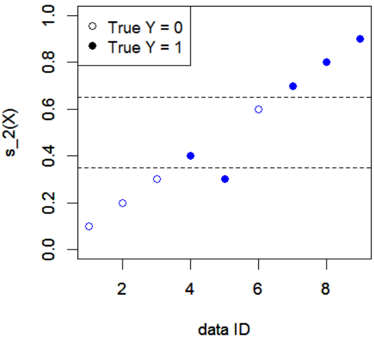
- Consider the following two classification functions $s_1(\mathbf{X})$ and $s_2(\mathbf{X})$:

ID	$s_1(\mathbf{X})$	$s_2(\mathbf{X})$	True state Y
1	0.4	0.1	0
2	0.4	0.2	0
3	0.4	0.3	0
4	0.4	0.4	1
5	0.5	0.3	1
6	0.6	0.6	0
7	0.6	0.7	1
8	0.6	0.8	1
9	0.6	0.9	1

- We can observe that when setting $c = 0.5$ for s_1 and $c = 0.4$ for s_2, both classifiers achieve the same highest accuracy of 77.8% (7/9).

(cont.) However, the classification function s_2 is more "robust" to the choice of c than s_1.





ROC curve and AUC

- Consider the following example:

ID	$s(\mathbf{X})$	True state Y
1	0.8	1
2	0.2	0
3	0.6	0
4	0.5	1
5	0.6	1

- Let c be a general threshold value such that $\widehat{Y} = \mathbf{1}\{s(\mathbf{X}) \geq c\}$.
- In this example, when $c = 0.8$,

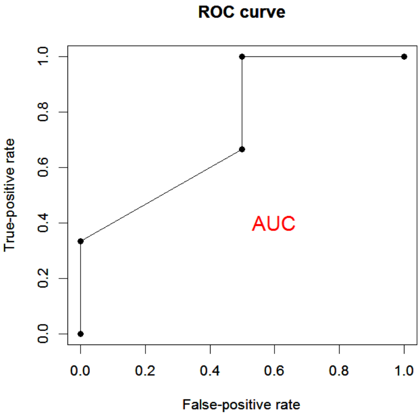
$$\Pr(\text{TP}|Y = 1) = 0.33(1/3), \;\; \Pr(\text{FP}|Y = 0) = 0(0/2).$$

(cont.)

- Similarly,
  - $c = 0.6 \Rightarrow \Pr(\mathsf{TP}|Y = 1) = 0.66(2/3)$, $\Pr(\mathsf{FP}|Y = 0) = 0.5(1/2)$.
  - $c = 0.5 \Rightarrow \Pr(\mathsf{TP}|Y = 1) = 1(3/3)$, $\Pr(\mathsf{FP}|Y = 0) = 0.5(1/2)$.
  - $c = 0.2 \Rightarrow \Pr(\mathsf{TP}|Y = 1) = 1(3/3)$, $\Pr(\mathsf{FP}|Y = 0) = 1(2/2)$.
- A graph of the true-positive rate $(\Pr(\mathsf{TP}|Y = 1))$ plotted against the false-positive rate $(\Pr(\mathsf{FP}|Y = 0))$ is called the ROC (receiver operating characteristic) curve.

ROC curve and AUC

- For the above example, we obtain the following ROC curve:



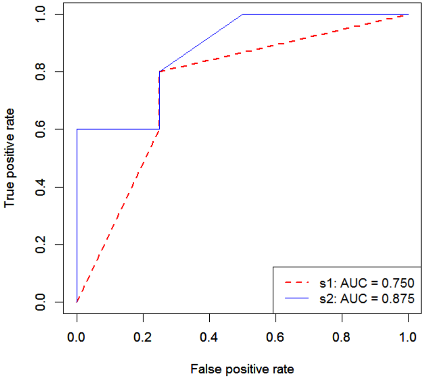
- The area under the ROC curve is referred to as AUC.

ROC curve and AUC

- AUC measures the precision of the classifier: larger AUC value indicates a better performance.
  - AUC $= 1$ corresponds to perfectly correct classification (in this case, there exists a threshold c^* such that the true-positive rate reaches to 1 while the false positive rate remains 0).
  - The worst possible case is AUC $= 0$.
  - Note that a pure random guess corresponds to the case of AUC $= 0.5$ (this is the case where the rate of true positive coincides with that of false positive).

ROC curve and AUC

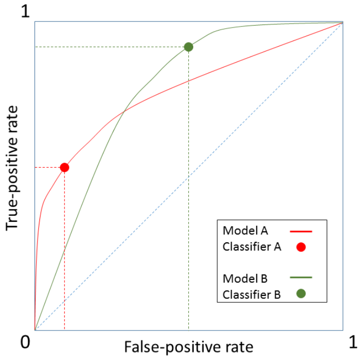
The ROC curves and AUC scores for the two classification functions given in p.13 are as follows:



Thus, in terms of AUC, s_2 is better than s_1.

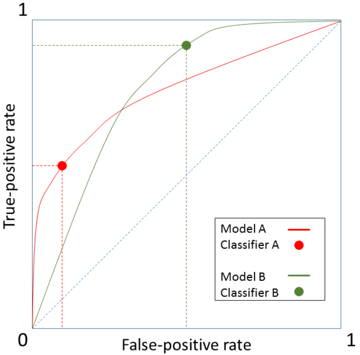
ROC curve and AUC

- It is important to note that AUC is not always a good measure of model performance, depending on your purpose.
- In the figure below, Model A and Model B have the same AUC scores.



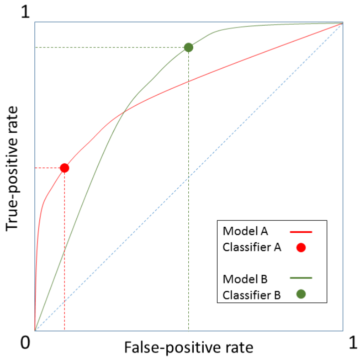
ROC curve and AUC

- Consider a cancer detection test, where achieving a high true-positive rate (the rate of detecting cancer in people who actually have it) is vital.
- Then, Classifier B (tpr $= 0.9$, fpr $= 0.5$) clearly outperforms Classifier A (tpr $= 0.5$, fpr $= 0.1$) .



ROC curve and AUC

- For another example, consider a spam email detection, where lowering the false-positive rate (the rate of misclassifying a regular email as spam) is more important.
- Then, in this case, Classifier A would be more useful than Classifier B.



Other measures

- There are many evaluation measures other than Accuracy and AUC. For example,
  - Type I error $= \frac{\#\text{FP}}{\#\text{FP}+\#\text{TN}}$
  - Type II error $= \frac{\#\text{FN}}{\#\text{TP}+\#\text{FN}}$

    Cancer detection: lowering Type II error is crucial.
    Spam mail detection: reducing Type I error is a primary concern.
  - Correlation coefficient (phi coefficient) between Y^{test} and $\widehat{Y}^{\text{test}}$.
- There is no "perfect" evaluation measure. We have to choose "right" measure(s) depending on the context.

Classification by Regression

- Practice datasets: training data **Titanic_train.csv**, test data **Titanic_test.csv**.
    - Data on passengers who were aboard the Titanic when it struck the iceberg on April 15, 1912.
    - This data is taken from the **Kaggle** website:[3]

      `https://www.kaggle.com/c/titanic`

- The data csv files are available from my website or from **Course Navi**.

[3] Kaggle is an online platform for data science competitions.

Classification by Regression

- In this exercise, we use **ROCR** package. Run the following code:

```
install.packages("ROCR")
library(ROCR)
```

- Set your working directory appropriately, and import the csv files:

```
setwd("C:/Rdataset")
train <- read.csv("Titanic_train.csv")
test  <- read.csv("Titanic_test.csv")
```

```
>  #install.packages("ROCR")
>  library(ROCR)
>
>  train <- read.csv("Titanic_train.csv")
>  test  <- read.csv("Titanic_test.csv")
>
>  head(train)
                             Passenger Survived Pclass    Sex Age SibSp Parch    Fare Embarked
1               Turja, Miss. Anna Sofia        1      3 female  18     0     0  9.8417        S
2         Francatelli, Miss. Laura Mabel       1      1 female  30     0     0 56.9292        C
3 Bishop, Mrs. Dickinson H (Helen Walton)      1      1 female  19     1     0 91.0792        C
4            Crosby, Capt. Edward Gifford      0      1   male  70     1     1 71.0000        S
5          Coelho, Mr. Domingos Fernandeo      0      3   male  20     0     0  7.0500        S
6                  Hunt, Mr. George Henry      0      2   male  33     0     0 12.2750        S
>  dim(train)
[1] 464   9
>  dim(test)
[1] 250   9
```

- The training data include 464 passengers, and the test data include 250.
- Using the training data, we build a classifier that predicts the survival status of the 250 passengers in the test data.

Classification by Regression

Definitions of variables

Survived	1 = Yes, 0 = No
Pclass	ticket class: 1 = 1st, 2 = 2nd, 3 = 3rd
Sex	male / female
Age	age in years
SibSp	the number of siblings / spouses aboard the Titanic
Parch	the number of parents / children aboard the Titanic
Fare	passenger fare
Embarked	port of embarkation: C = Cherbourg, Q = Queenstown, S = Southampton

Classification by Regression

- We first transform the non-numeric variables into numerical variables.

```
Male <- train$Sex == "male"
EmbC <- train$Embarked == "C"
EmbQ <- train$Embarked == "Q"
```

- "Southampton" is not used as the benchmark (to avoid linear dependence).

Classification by Regression

- Linear regression (linear probability model):

```
lin <- lm(Survived ~ Pclass + Male + Age + SibSp + Parch + Fare +
       EmbC + EmbQ, data = train)
summary(lin)
```

- Logistic regression (binary logit model):

```
log <- glm(Survived ~ Pclass + Male + Age + SibSp + Parch + Fare +
       EmbC + EmbQ, data = train, family = binomial(link = "logit"))
summary(log)
```

```
>  summary(lin)

Call:
lm(formula = Survived ~ Pclass + Male + Age + SibSp + Parch +
    Fare + EmbC + EmbQ, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05447 -0.22437 -0.06577  0.21374  0.98634

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2892162  0.0976365  13.204  < 2e-16 ***
Pclass      -0.1679857  0.0285876  -5.876 8.12e-09 ***
MaleTRUE    -0.5051579  0.0382026 -13.223  < 2e-16 ***
Age         -0.0061591  0.0013319  -4.624 4.90e-06 ***
SibSp       -0.0691233  0.0208240  -3.319 0.000975 ***
Parch       -0.0362843  0.0223862  -1.621 0.105746
Fare         0.0005757  0.0004556   1.264 0.206989
EmbCTRUE     0.0328273  0.0482599   0.680 0.496712
EmbQTRUE    -0.0430323  0.0847041  -0.508 0.611678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>  summary(log)

Call:
glm(formula = Survived ~ Pclass + Male + Age + SibSp + Parch +
    Fare + EmbC + EmbQ, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7100  -0.5966  -0.3625   0.5851   2.4752

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.277866   0.799771   6.599 4.13e-11 ***
Pclass      -1.151116   0.216059  -5.328 9.94e-08 ***
MaleTRUE    -2.838642   0.284862  -9.965  < 2e-16 ***
Age         -0.044956   0.010482  -4.289 1.79e-05 ***
SibSp       -0.566009   0.174055  -3.252  0.00115 **
Parch       -0.215026   0.166733  -1.290  0.19718
Fare         0.003996   0.003654   1.093  0.27420
EmbCTRUE     0.174492   0.345783   0.505  0.61382
EmbQTRUE    -0.728700   0.710058  -1.026  0.30477
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction on the test data

- Create $(\mathbf{X}^{\text{test}}, Y^{\text{test}})$:

```
testY <- test$Survived
testX <- with(test, cbind(1, Pclass, Sex == "male", Age, SibSp,
       Parch, Fare, Embarked == "C", Embarked == "Q"))
```

- Computation of $s(\mathbf{X}^{\text{test}})$:

```
s.lin <- testX%*%lin$coef
s.log <- testX%*%log$coef
```

Classification by Regression

- Create an object of class "prediction":

```
pred.lin <- prediction(s.lin, testY)
```

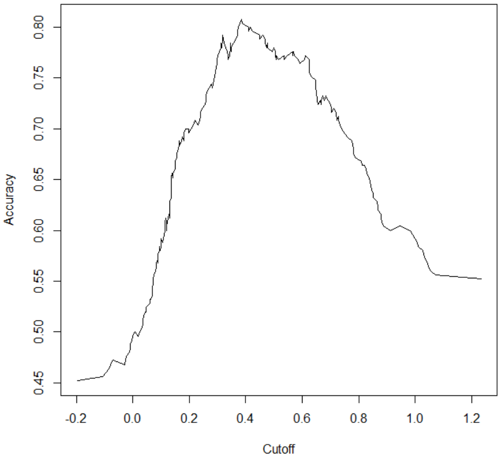
- Accuracy, AUC and tpr-fpr:

```
acc.lin <- performance(pred.lin, "acc")
auc.lin <- performance(pred.lin, "auc")
roc.lin <- performance(pred.lin, "tpr", "fpr")
```

- Plot of cut-off vs. accuracy:

```
plot(acc.lin)
```

Classification by Regression



Classification by Regression

```
>  max(acc.lin@y.values[[1]])
[1] 0.808
>  which.max(acc.lin@y.values[[1]])
[1] 100
>  acc.lin@x.values[[1]][100]
[1] 0.383763
```

- The linear regression classifier achieves the highest accuracy of 80.8% when $c = 0.384$.

Classification by Regression

- AUC score:

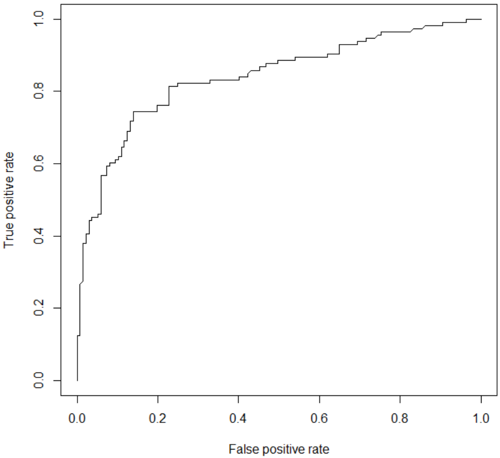
```
auc.lin@y.values[[1]]
```

```
> auc.lin@y.values[[1]]
[1] 0.8400297
```

- Plot of ROC curve:

```
plot(roc.lin)
```

Classification by Regression



Classification by Regression

- For the logistic regression classifier, similarly as above, define

```
pred.log <- prediction(s.log, testY)
acc.log  <- performance(pred.log, "acc")
auc.log  <- performance(pred.log, "auc")
roc.log  <- performance(pred.log, "tpr", "fpr")
```

- Following the same procedure as above, we can find that the highest accuracy is 81.2% at $c = -0.016$, and the AUC score is about 0.838.
- Thus, the results show that both classifiers have almost the same prediction performance.[4]

[4] Recall that in the econometric analysis of binary responses, using a linear regression is problematic in terms of "interpretation". However, for the purpose of "prediction", using a linear model is not a problem at all .