

Applied AI - Classification

# Week	2
▼ Course	DES423 - Applied ML & AI
▼ Type	Lecture
🕒 Created	@August 18, 2022 9:04 AM
☑ Reviewed	<input type="checkbox"/>

Applied Machine Learning → Understand it at the abstract level

Classification

Type: Binary, Multi-class, Multi-label Classification

Binary Classification

classify into two classes: true or false ex. spam or not spam

Multi-class Classification

classify into more than 2 classes ex. face classification, plant species, OCR

Multi-label Classification

like multi-class but more than 1 class can be predicted


Example: Iris Classification

Classical Example of Classification Problem

3 Classes: Iris Sentosa, Versicolor, Virginica

Used 2 features: Petal and Sepal

Google Colaboratory

 <https://colab.research.google.com/drive/1BbMwRf8xPMLAH1lv-JvXipajb-XPnRYz?usp=sharing>



Basic ML Process [Recap]

Dataset —Split→ Training Set (to train algorithm) and Test Set (to test the algorithm)

x_{train} , y_{train} used to train the model

x_{test} , y_{test} used to test and evaluate model → yield $y_{predict}$ and compare with y_{test} (actual value)
→ obtain accuracy score.

KNN - K-Nearest Neighbors Algorithm

Which class is the nearest to the unknown data, that unknown data likely to belong to that class.

KNN is Supervised ML

Classifies data point based on how neighbors are classified.

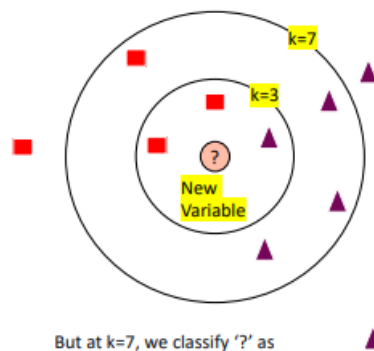
K = number of nearest data we want to compare with

Used the majority vote on what unknown data point is likely to be.

How to choose the K

choosing the K value → **Parameter Tuning** for better accuracy

The small number, the higher noise influence . The larger number, more expensive calculation.



To choose the value of K,

- \sqrt{n} where n is the total number of data points
- Odd value of K is selected to avoid confusion and tie-breaking.

We will use KNN when the data should be **labeled**. If there is 0 or null, you need to label it or something else. Data should be **noise-free**. The dataset is **small**.

KNN is a lazy learner, they just collect the data.

KNN calculate the distance using Euclidean distance (distance between two points)

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Calculate the distance from the unknown point to each point in the dataset
- Sort the distance from low to high
- Select the best k result and look at its class.
- Which class has the most, the unknown data point belongs to that class.

Overfitting

This is when training set does good e.g. like accuracy at 1.0 but failed on test set. You need to adjust the parameter or train using another algorithm.

Confusion Matrix

Column as Predicted Value, Row as Ground Truth Value

Example: DB-Diabetes and Non-DB

GT/Predicted	Non-DB	DB
Non-DB	94	13
DB	15	32

Accuracy

$Accuracy = \text{correct prediction} / \text{total}$

Solution: $(94+32) / (94+32+15+13) = 0.81$

Assignment 1

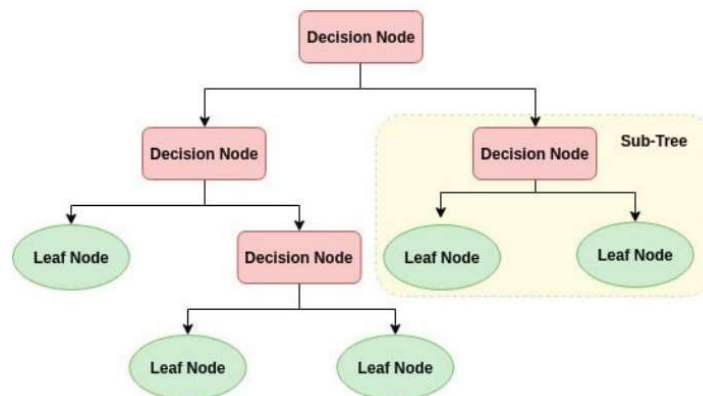
<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/6de5e405-62e0-43af-ab73-03311c054bbe/DES423-Assignment-1-6222780379.pdf>

Applied AI - Decision Tree

# Week	3
Course	DES423 - Applied ML & AI
Type	Lecture
Created	@August 25, 2022 8:51 AM
Reviewed	<input type="checkbox"/>

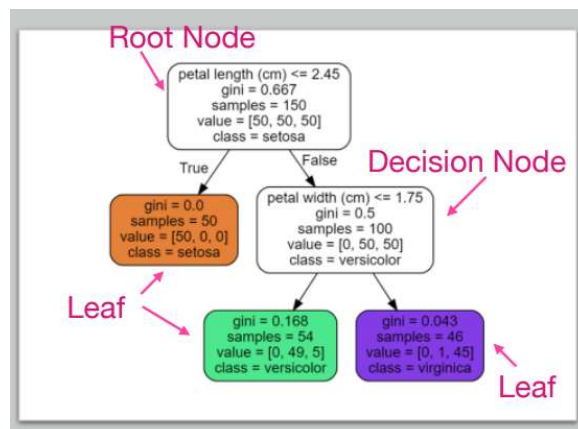
Decision Tree

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/d2a91407-879f-4207-adb1-3bece0cfedf7/DES423-Lecture-03-Decision-Tree.pdf>



All the node present the decision in the tree, leaf node is a label.

Decision tree is the white-box algorithm which you can see how its work and easy to explain it.

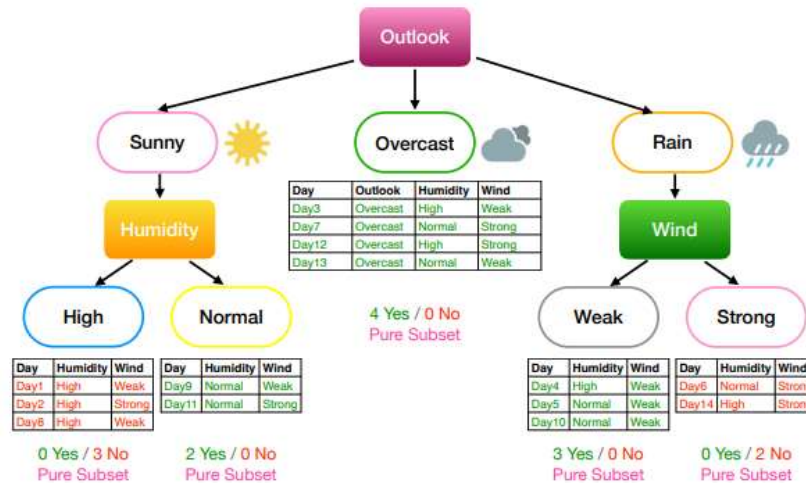


At the decision node, it usually present the feature of the dataset.

Example: Will he come to work?

Try to understand when he works. We are going to make decision tree based on his history.

Steps: split into subset, then check if the set of result is pure or not, if yes stop, if no then repeat the process.



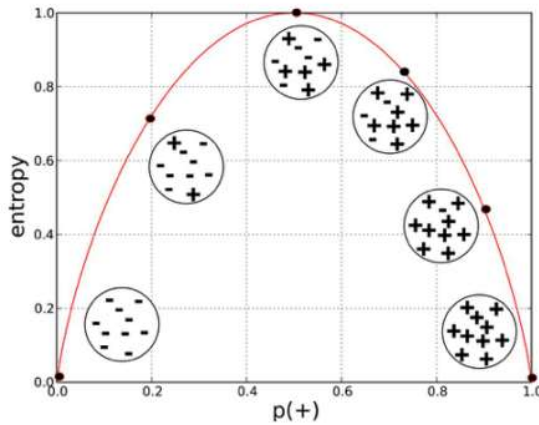
Sometime we have to decide which feature should be the first one to be decided in the tree.

ID 3 Algorithm

- Split (node, {examples}):
 - A <- the best attribute for splitting the {examples}
 - Decision attribute for this node <- A
 - For each value of A, create new child node
 - Split training (examples) to child nodes
 - For each child node / subset:
 - If subset is pure: STOP
 - Else: Split (child_node, {subset})

Entropy

Entropy is the measure of the **uncertainty** in the group of observations. It used to determine how to split the data in the decision tree.



It started at 0 when there is all certain outcome to be the same, then it is increasing until 1 when all outcome is evenly divided (50/50).

Entropy is in the range of 0 to 1, inclusive.

Entropy is 1 when the probability is 50/50. Entropy is 0 is there is a **certainty** of the outcome.

$$Entropy(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

Where S is the collection of sample, p+ is the portion of positive sample, p- is the negative portion.

Example of calculation

$$\begin{aligned} &\text{when there are 5+ , 15-} \\ &p_+ = 5/20 = 0.25; p_- = 15/20 = 0.75 \\ &Entropy([5+, 15-]) = -(0.25)(\log_2(0.25)) - (0.75)(\log_2(0.75)) \\ &= 0.36 \end{aligned}$$

Information Gain

Information gain is the expected reduction in entropy caused by dividing the example of the sample according to attribute.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- More precisely, the information gain, **Gain(S, A)** of an attribute **A**, relative to a collection of example **S**, is defined as,

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where **Values(A)** is the set of all possible values for attribute **A**, and **S_v**, is the subset of **S** for which attribute **A** have value **v**.

To choose the attribute for the first decision, choose the one that has the maximum information gain.

Gini Index

The Gini impurity is in 0 to 0.5

General calculation for Gini Index; $Gini(S) = 1 - \sum_j p_j^2$

If there is 2 classes; $Gini(S) = 1 - (p_+^2 + p_-^2)$

Gini is 0 when the node is pure; Gini is faster than Entropy because the entropy calculation has the logarithm.

Choose the higher Gini Gain attributes to be the decision node.

The decision tree model is fit very well with the train dataset. But, it will give the wrong result when used to predict the value.

Advantages and Disadvantages

- Simple to understand and interpret, require a little data preparation, can handle both numerical and categorical data.
- However, it can create over-complex tree → Overfitting, unstable tree (sensitive to data, if data change, tree can change), can create biased tree when some classes dominated.

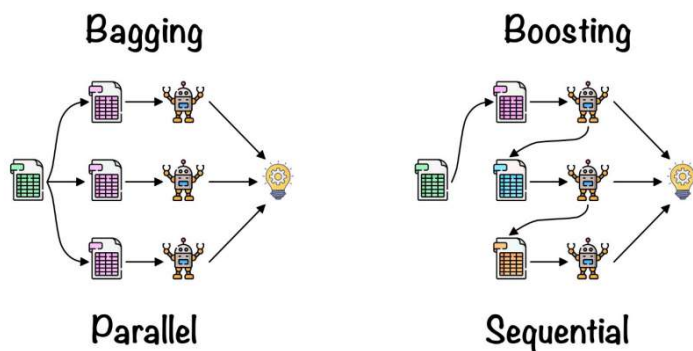
Random forest will help with the overfitting problem of the decision tree.

Random Forest: Ensemble learning method.

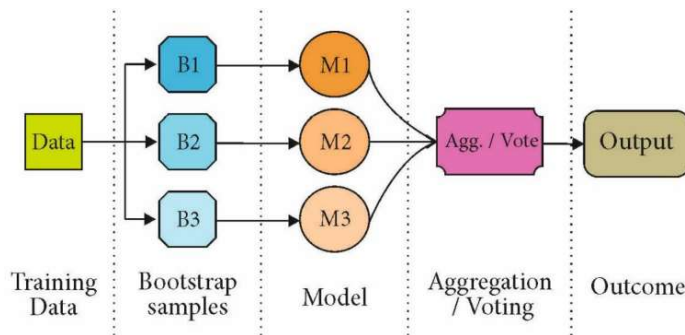
Bagging and Boosting

Ensemble learning method used to reduce the overfitting and improve the variant of the model to improve the accuracy.

In this class, bagging method will be implemented.



Bootstrap Aggregating



Random forest is common in ML which combines the output of many decision tree to reach a single result.

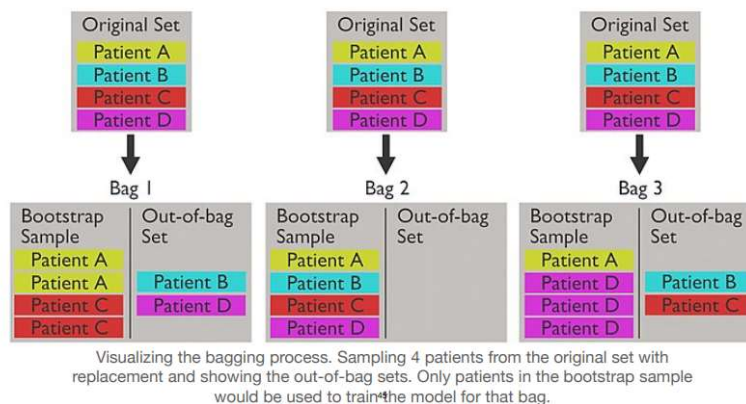
Bagging introduce randomness into the rows of the data, but random forest introduces to the rows and columns of data. Combined, provides more diverse decision tree set and decrease prediction errors.

Bootstrapped dataset → Create many dataset to have different data from the original; some bootstrapped dataset an have the same row of data in it. (Bootstrapped → Bagging concept)

Random forest we don't need to use all the features, pick the subset of features → create variety of decision trees. (Original bagging you need to use all attribute) **We need uncorrelated decision tree model.**

Out-of-bag Error (OOB)

Use out of bag set to test the tree in the random forest.



Advantages and Disadvantages of Random Forest

- Less prone to overfitting, highly accurate and robust method (If we have the missing data, this model still can predict the data), high dimensional data, can give the important feature after training.
- However, it is a black box, computation may be more complex to other algorithm.

Applied AI - Naïve Bayes

# Week	4
▼ Course	DES423 - Applied ML & AI
▼ Type	Lecture
🕒 Created	@September 1, 2022 9:27 AM
☑ Reviewed	<input type="checkbox"/>

Lecture file:

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/fa7b8e0a-58e7-4233-b761-1d100fae65e4/DES423-Lecture-04-Naive-Bayes-Revised.pdf>

Naïve Bayes

Mostly used in spam email classification and other text classification.

Naïve Bayes is based on the concept of conditional probability.

If it is








$P(A|B)$ is the probability of A given B, in case of dependent event.

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

$P(A)$ is the probability that A is happening, $P(B)$ is the probability that B is happening, $P(A \cap B)$ is the probability of both A and B happening.

Noted that $P(A|B) \neq P(B|A)$, it may in some case but mostly it isn't.

Monty Hall Problem

 1/3	The car has 1/3 chance behind any door  1/3		 1/3
The host has a 100% chance eliminating the door with the goat behind	The host has 1/2 chance eliminating any door when 2 of them have a goat behind		The host has a 100% chance eliminating the door with the goat behind
 100% Eliminate Door 1	 1/2 Eliminate Door 1	 1/2 Eliminate Door 2	 100% Eliminate Door 2
Switch & got Car 1/3	Switch & got Goat 1/6	Switch & got Goat 1/6	Switch & got Car 1/3

Bayes Theorem

It gives us the conditional probability of event A given the event B has already occurred.

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

$$P(A|B)P(B) = P(A \cap B)$$

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

$P(A|B)$ → Posterior Probability

$P(B|A)$ → Likelihood, $P(A)$ → Prior Probability, $P(B)$ → Evidence

Where is Naïve Bayes used?

Usually used in face recognition, weather prediction, medical diagnosis, news classification → classify the category of news should be in.

Naïve Bayes Classifier

From shopping example, predict that a person will purchase a product on a specific combination of day, discount, and free delivery or not using Naïve Bayes classifier.

Frequency Table

Frequency Table		Purchase	
		Buy	Not Buy
Discount	Yes	19	1
	No	5	5

Based on the dataset containing the three input types—day, discount, and free delivery—the frequency table for each attribute is populated.

Frequency Table		Purchase	
		Buy	Not Buy
Day	Weekday	9	2
	Weekend	7	1
	Holiday	8	3

Frequency Table		Purchase	
		Buy	Not Buy
Free Delivery	Yes	21	2
	No	3	4

We can use this information to predict whether the customer will buy or not buy the product.

Later, create a likelihood table

Frequency Table		Purchase(A)		
		Buy	Not Buy	
Day(B)	Weekday	9	2	11
	Weekend	7	1	8
	Holiday	8	3	11
		24	6	30

Frequency Table		Purchase(A)		
		Buy	Not Buy	
Day(B)	Weekday	$P(\text{Weekday} \text{Buy})$	$P(\text{Weekday} \text{Not Buy})$	$P(\text{Weekday})$
	Weekend	$P(\text{Weekend} \text{Buy})$	$P(\text{Weekend} \text{Not Buy})$	$P(\text{Weekend})$
	Holiday	$P(\text{Holiday} \text{Buy})$	$P(\text{Holiday} \text{Not Buy})$	$P(\text{Holiday})$
		$P(\text{Buy})$	$P(\text{Not Buy})$	

Frequency Table		Purchase(A)		
		Buy	Not Buy	
Day(B)	Weekday	9/24	2/6	11/30
	Weekend	7/24	1/6	8/30
	Holiday	8/24	3/6	11/30
		24/30	6/30	

Then, calculate using the Bayes theorem. (You can calculate only the top part, not the denominator part because it is constant so you can ignore it)

Shopping Example - Not Buy

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

Frequency Table		Purchase		
		Buy	Not Buy(A)	
Day	Weekday	2/5	2/5	4/10
	Weekend	2/5	1/5	3/10
	Holiday(B)	1/5	2/5	3/10
		5/10	5/10	

Frequency Table		Purchase		
		Buy	Not Buy(A)	
Discount	Yes(B)	4/5	1/5	5/10
	No	1/5	4/5	5/10
		5/10	5/10	

Frequency Table		Purchase		
		Buy	Not Buy(A)	
Free Delivery	Yes(B)	3/5	2/5	5/10
	No	2/5	3/5	5/10
		5/10	5/10	

A = Not Buy

B = Discount:Y, Delivery:Y, Day=Holiday

Applying Bayes Theorem, we get $P(A|B)$ as shown:

$P(A|B) = P(\text{Not Buy} | \text{Discount:Y, Delivery:Y, Day=Holiday})$

$$= \frac{P(\text{Discount : Y, Delivery : Y, Day = Holiday} | \text{Not Buy}) * P(\text{Not Buy})}{P(\text{Discount : Y, Delivery : Y, Day = Holiday})}$$

$$= \frac{P(\text{Discount : Y} | \text{Not Buy}) * P(\text{Delivery : Y} | \text{Not Buy}) * P(\text{Day = Holiday} | \text{Not Buy}) * P(\text{Not Buy})}{P(\text{Discount : Y}) * P(\text{Delivery : Y}) * P(\text{Day = Holiday})}$$

$$= \frac{\frac{1}{5} * \frac{2}{5} * \frac{2}{5} * \frac{5}{10}}{\frac{5}{10} * \frac{5}{10} * \frac{3}{10}}$$

$$= 0.213$$

Assignments

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/5ba944ff-865b-4f42-8248-e31d30f02ba5/DES423-Class-04-Assignment-01.pdf>

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/8c83a140-ee95-42d6-b909-df40588043f6/DES423-Class-04-Assignment-02.pdf>

Applied AI - SVM

# Week	4
▼ Course	DES423 - Applied ML & AI
▼ Type	Lecture
🕒 Created	@September 1, 2022 11:21 AM
☑ Reviewed	<input type="checkbox"/>

Lecture file:

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/1313e387-bb4e-41a3-b35a-1fa0a0d3d902/DES423-Lecture-05-SVM.pdf>

Support Vector Machine

Divide dataset using simple line (linear) or non-linear line that best separate the dataset with the biggest gap of each group separation as much as possible.

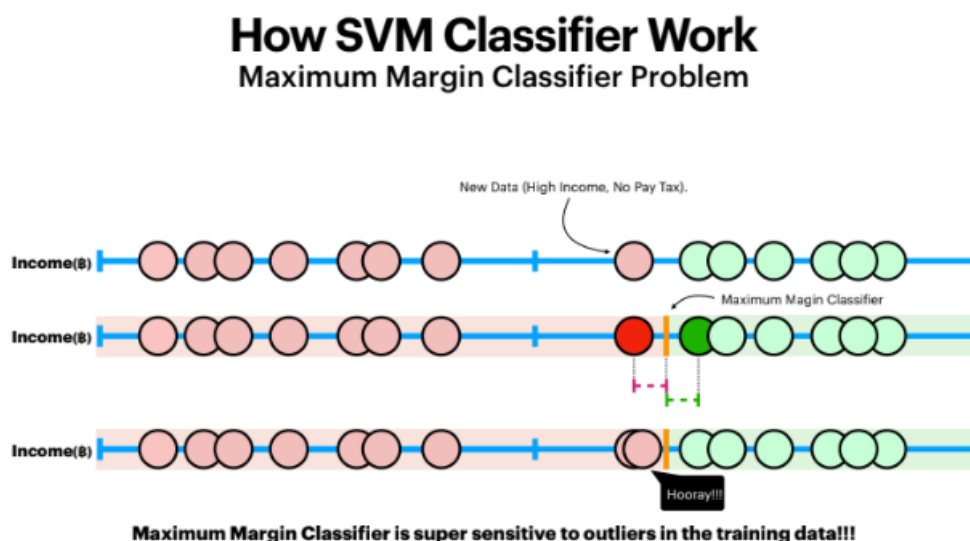
SVM can be used for both regression and classification tasks. (But first we focused on classification first)

How SVM works?

Basically record the data first, then label the data

However, If we use the maximum margin classifier for determining where we will separate the data, it is sensitive to outliers.

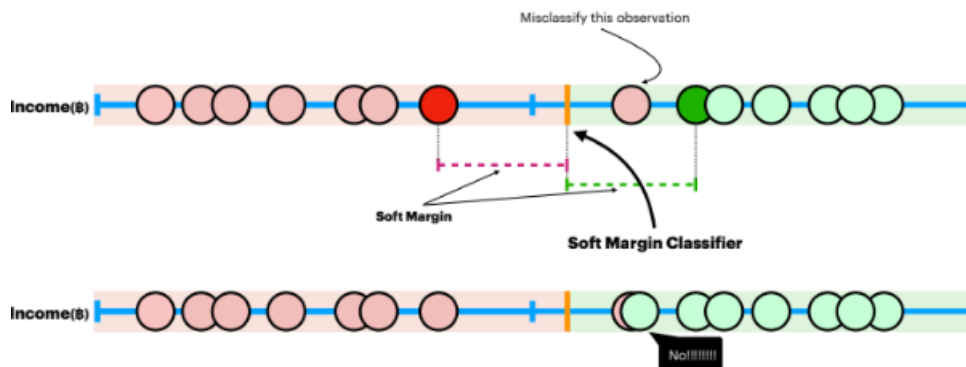
The line that gives the maximum margin is the best separator or classifier in the SVM. If the data point is on the margin line, it is called the support vector, data fallen in between the margin line is the misclassified data point.



There is **Soft margin classifier**. If there is an outlier, we still have the correct prediction for new data.

How SVM Classifier Work

Soft Margin Classifier



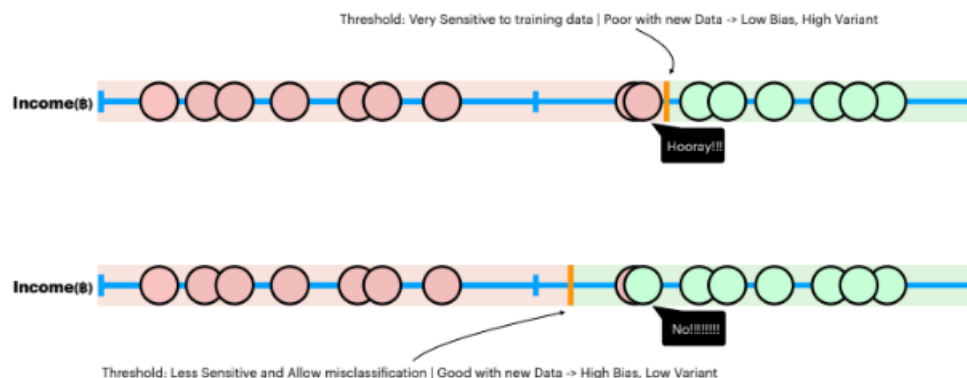
For cross validation, we need to find the best soft margin position. We use cross validation to determine how many misclassification and observation to allow inside the soft margin to get the best classification and verify the model to suit the problem the most and give the best possible model.

Bias and Variance Tradeoff

It is stick too good with the data. Low bias → High variant, but High bias → Low variant.

How SVM Classifier Work

Bias/Variance Tradeoff



2D SVC

Hyperplane

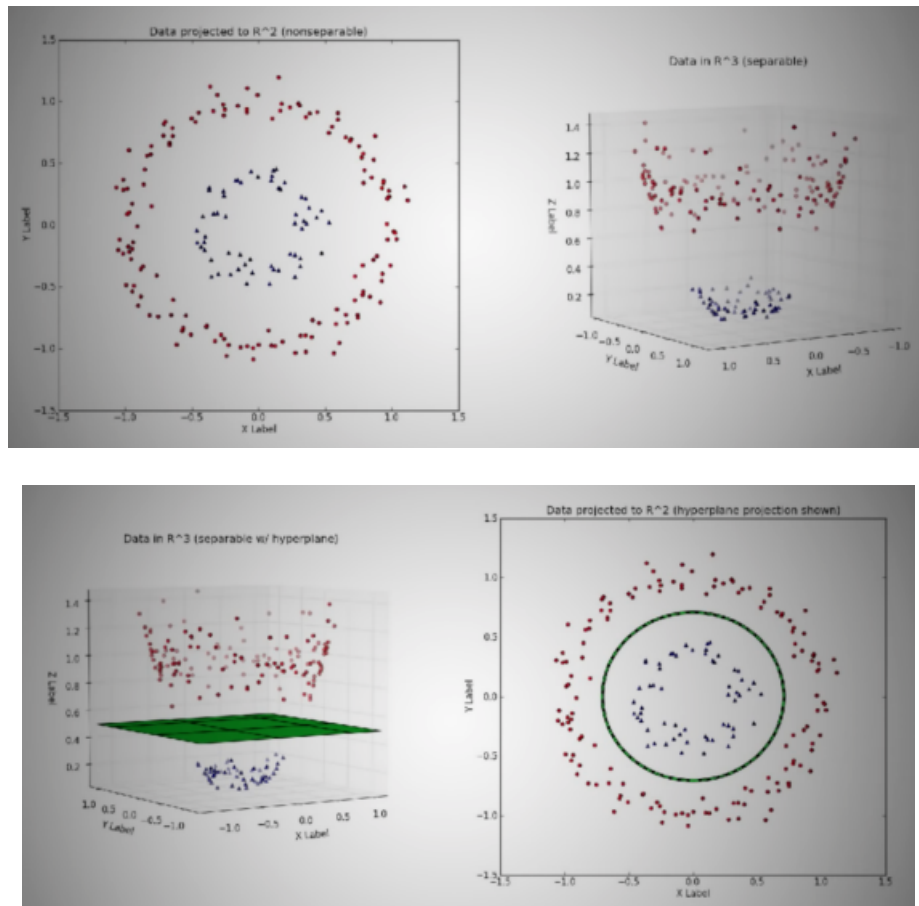
The one that split data into one less dimensions for easier classification. ($n-1$ dimension plane in n dimension data)

- 1D → Point hyperplane

- 2D → Line hyperplane
- 3D → Plane hyperplane

Non-linear SVC

From 1D data, you transform into 2D by applied some function for easier separation of the classification.
(Basically increase the dimension of the data and find the hyperplane from that)

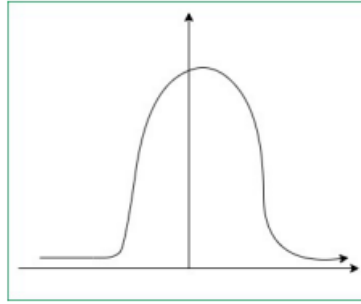


This is called “**Kernel function**”

Kernel function

Used to take data and transform into the required form to separate data. Kernel is usually a set of mathematical function to manipulate the data. For example,

- Polynomial kernel
- Sigmoid Kernel
- Gaussian Kernel Radial Basis Function (RBF)



RBF Function graph

Kernel Trick

Allowed to use original dataset without computing the transformed higher dimensional dataset. It just look at the relation or something to the data. It is the more efficient and less computationally expensive to transform data into higher dimension.

Advantages and Disadvantages

- The real strength of SVM, less risk from overfitting, very effective with high dimension data.
- However, not effective with large dataset, full labeling of input data is required, choose optimal kernel for SVM (difficult task), hyper-parameter is also hard to fine-tune.

Noted that for the classification problem, it is based on the fundamental problem of the task. It can be varies from problem to problem.

Applied AI - Regression

# Week	5
▼ Course	DES423 - Applied ML & AI
▼ Type	Lecture
🕒 Created	@September 8, 2022 8:58 AM
☑ Reviewed	<input type="checkbox"/>

Lecture file

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/d1dae281-4b3a-4b21-b7c4-2b100fab402b/DES423-Lecture-06-Regression.pdf>

Table of Content

Regression

Applications of regression

Linear Regression

Equations

Sum of Squared Errors

Polynomial Regression

Equation of polynomial

Advantages and Disadvantages

Multiple linear regression

Regression

“Try to predict the continuous data”

Applications of regression

For example, economic growth to predict the GDP in the future, predict the price of the product in the future and housing price, score prediction. It is depends on the complexity and continuity of data in the problem.

Linear Regression

Statistical model to predict the relationship between independent and dependent variables by examining two factors:

- What are significant predictors of the outcome?
- How significant the regression line in terms of making prediction with the highest possible accuracy?

“Dependent variable” → One variable is affecting another variable. For example, rainfall and water level in the river.

“Independent variable” → Not change or affects another variable.

Use independent variable to predict the dependent variable

For example to determine the dependent/independent variables: House Price

Independent variable: Number of rooms, Location, Size

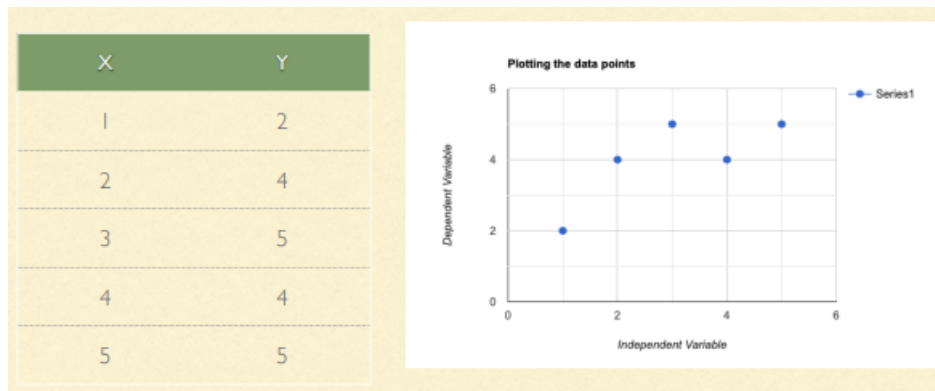
Dependent variable: Price of the house

Equations

Linear regression; use linear line as a prediction equation

Simple line equation: $y = mx + c$

How can we draw the regression line?



There are 2 ways to calculate the regression line.

1. Calculate the mean point and the regression line should pass the mean point. (Mean point is calculate by using mean of X and mean of Y value).
2. Use complicated equation and replace in the line equation.

$$Y = \beta X + \alpha$$

$$\beta = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

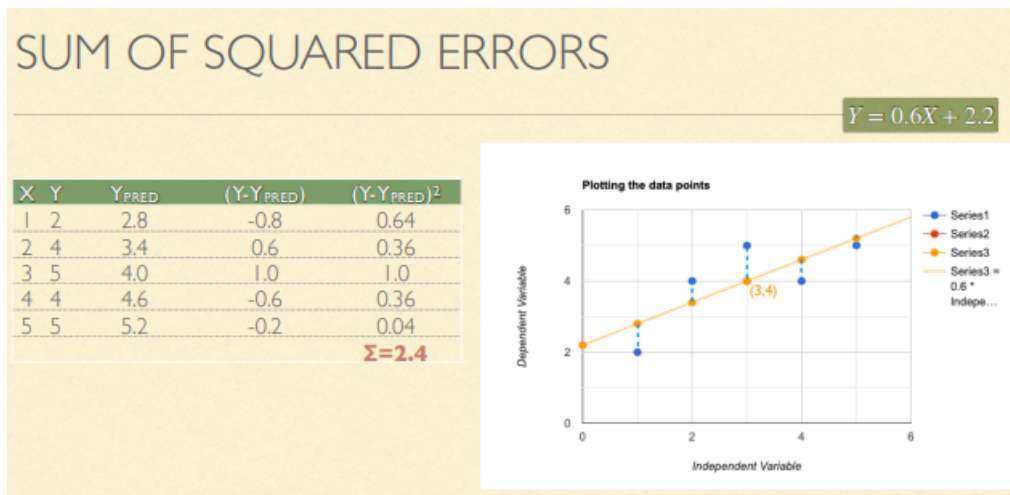
$$\alpha = \bar{Y} - \beta \bar{X}$$

If you use $y=mx+c$, here is the faster equation for you.

$$m = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$c = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

Sum of Squared Errors



You can use gradient descent to calculate for least SSE. (MSE problem in the scientific computing class).

Assignment 01

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/a5782a96-21ef-4538-a808-ef9279d6fc9b2/DES423-Class-05-Assignment-01.pdf>

Polynomial Regression

Simple linear regression only works when relationship of the data is linear. Polynomial regression helps to identify the relationship between variables in non-linear way.

Equation of polynomial

a_0 is the intercept, a_1 - a_n are coefficient. (I assume)

$$y = a_0 + a_1x_1 + a_2x_2^2 + a_3x_3^3 + \dots + a_nx_n^{n-1}$$

- Degree of order to use is a **hyperparameter**
- A higher degree tries to **overfit** (fit too well with the training data), lower degree tries to **underfit** (not fit with the training data).
- Estimate the relationship between coefficients → it is still the linear regression so mostly it is called **polynomial linear regression**.

random state in the `train_test_split` is to randomize train dataset every time you train the model to find the more accurate model.

Advantages and Disadvantages

- Polynomial regression provide best approximation of the relationship between the dependent and independent variable.
- If there are outliers, it can affect the overall result of the model.

Multiple linear regression

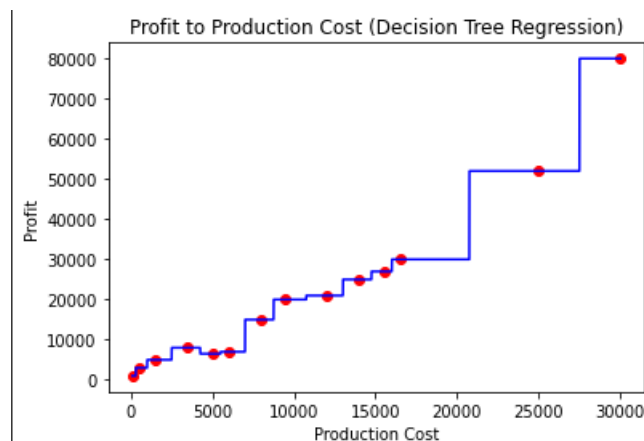
It is similar to polynomial regression, but instead of power of variable, it is use another variable.

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n$$

For example, the advertising data, how much money you put in different medium will generate how many sales units. So, in this case, a_0 is still the intercept, a_1 to a_n is the coefficient, but x_1 to x_n is the variable represent the amount of money you spend in the advertising medium.

Noted that there is still many type of regression such as tree regression, logistics regression, support vector regression, and lasso regression etc.

The **decision tree regression** graph line is the result of regression, it can look like a step graph. To generate the decision tree, instead of using information gain such as entropy and Gini index, they use sum of squared error for this case.



`StandardScaler` will use when you want to scale down the value but keep the meaning of the data. Usually use in the case that data is very big or large differences.

Assignment 2:

Which regression model is used for classification problem?

Answer:

Logistic regression can be used to perform the binary classification task.

Logistic regression is based on the concept of probability. The sigmoid function is used to return the probability of a class or a label, with the ability to map any input to output between the range of 0 to 1, inclusive. The threshold value can decide which class it belongs to.

Applied AI - Gradient Descent and Model Selection

# Week	6
Course	DES423 - Applied ML & AI
Type	Lecture
Created	@September 15, 2022 9:33 AM
Reviewed	<input type="checkbox"/>

Gradient Descent

Sum of square error

Gradient Descent core

Model Selection

Accuracy

Holdout sets

Cross-validation method

Regression → Bias-Variance trade-off

Validation curves

Model Evaluation

Confusion Matrix

Accuracy and Precision

Multiple Class confusion matrix

Assignment

Lecture file:

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/af42e3a0-ad80-48cf-9633-381238808146/DES423-Lecture-07-Gradient-Descent.pdf>

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/5cacf41a-3a4d-493f-afdd-34a9eeddafad/DES423-Lecture-07-Method-Selection.pdf>

Gradient Descent

Try to reduce or create a relationship between $y=ax+b$ if given the data (0,1) , (1,3) , (2,5) , (3,7)

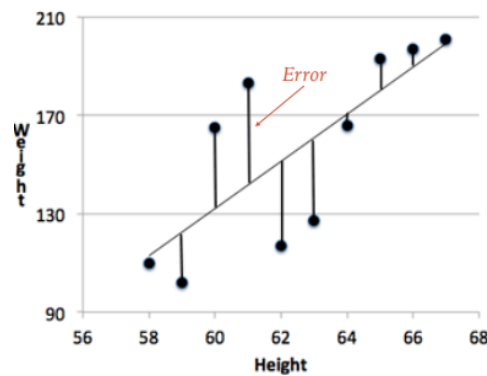
Human can think that it is $y=2x+1$, but how will computer do it? It guesses then find the error and adjusting it until finding nearest solution

The different between actual and predicted solution ($y-h$) is called error → it should be small to be good. Loss function can be calculated for calculator, Square error is the most widely used one.

Sum of square error

SSE → can be cost function

$$SSE = \sum_{i=1}^n \frac{1}{2} (y_i - h_i)^2$$



Gradient Descent core

A gradient measure how the output of a function changes if you change the input a little bit.

We use two equation to find the best way to get to the local minimum

1. Guess a and b
2. Calculate cost function (in this case, SSE)
3. Find slope along axis a and b; $a = \frac{\partial(SSE)}{\partial a}$ and $b = \frac{\partial(SSE)}{\partial b}$
4. Adjusting a and b according to the learning rate alpha (it should be very small);
 $a = a - \alpha \frac{\partial(SSE)}{\partial a}$ and $b = b - \alpha \frac{\partial(SSE)}{\partial b}$
5. Repeat 2-4 until get lower cost or cost $\gg 0$

This is just the concept of gradient descent.

Model Selection

How can we should selection of model for our problem? Most people will says accuracy.

Accuracy

The problem of this is that there are multiple way you can come up with the accuracy number.

If you use the same train and test set data as the same one, it may not reflex the true result.

So, you divide the train and test set

Holdout sets

In sklearn library, they have `train_test_split` function to split the dataset into 2 sets as train and test set to use that as a validation correctly. `train_size` is the size of train dataset (usually 70%).

However, there is no single number of accuracy of your model that you can tell others people. There are other ways to do this.

Cross-validation method

The very important method. The method that we use is K-fold cross validation.

If $k=2$, you split the dataset as normal but split into two set, and use one as train, then train again using another set. When validating the accuracy each time, use the remaining unused set.

If $k=5$, we will get 5 accuracy number, then average it out.

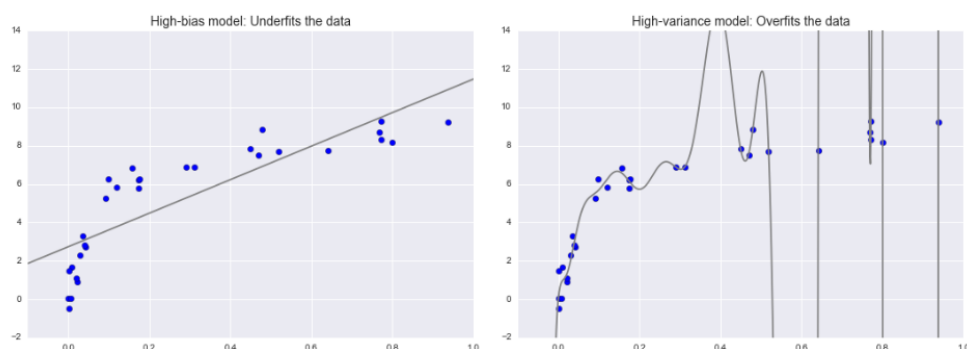
Round	Portion 1	Portion 2	Portion 3	Portion 4	Portion 5
1	Train set	Train set	Train set	Train set	Validating Set
2	Train set	Train set	Train set	Validating Set	Train set
3	Train set	Train set	Validating Set	Train set	Train set
4	Train set	Validating Set	Train set	Train set	Train set
5	Validating Set	Train set	Train set	Train set	Train set

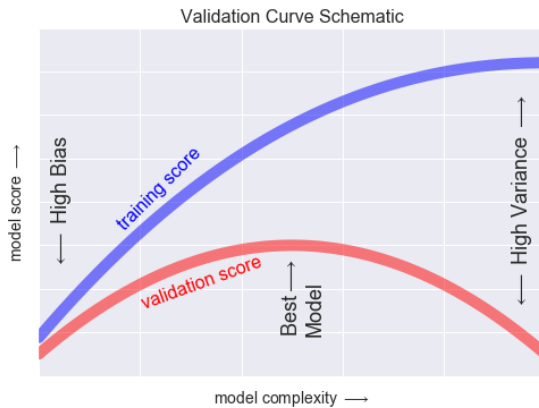
`cross_val_score` is the function in sklearn library to provide this method, you define the number of fold in the cv parameters; `cross_val_score(model, x, y, cv=5)`

If you decided to use **Leave one out method**, it will yield the cross validation score for all samples. Like leave the one record out to be the validation set, use the remaining as train set.

Regression → Bias-Variance trade-off

find the sweet spot of the trade off, use less polynomial → underfit, use more polynomial → overfit





The model will work well with training data but lower validation score if the model complexity is too high, it will also yield high variance.

Validation curves

will be used for polynomial regression, comparing different number of degree and find what is the best one to fit with our problem.

Model Evaluation

Accuracy is not the best one to tell the performance of the model. It will only works well if the data is balanced for all classes.

Confusion Matrix

It will tell you more than the accuracy of the model. It uses to evaluate the classifier and the classification problem.

Column as predicted value, row as actual value

Actual/Predicted	Predicted True	Predicted False
Actual True	True Positive: TP	False Negative: FN
Actual False	False Positive: FP	True Negative: TN

So, the metrics are

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Negative Predicted Value} = \frac{TN}{TN + FP}$$

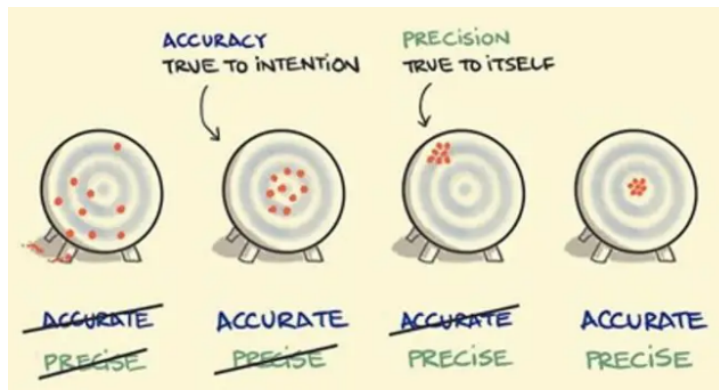
$$F_1 \text{ Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Precision think about the predicted positive value only, but recall using the actual positive value.

The number you should reduce in the spam email classification problem → you want to reduce the number of false positive (actual not, predicted yes) because that email will go unseen by the users. However, if you change the problem to cancer detection problem → you want to reduce the number of false positive error because you don't want people to have medication that they don't need.

If you have to compare between 2 or more models, you use F1 score to compare them.

Accuracy and Precision



Precision → Out of **all predicted true**, how many you got right? → Prediction as base

Recall → Out of **all actual true**, how many you got right? → Actual as base

Multiple Class confusion matrix

You need to know TP/TN/FP/FN of each class then calculate for each class. Use the average value from all classes to represent that number.

Assignment

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/a8144341-dcfe-4005-ab95-6ba3c48e1e49/DES423-Class-06-Assignment-01-6222780379.pdf>