應用機器學習

Brian Chan 陳醒凡

課程目標

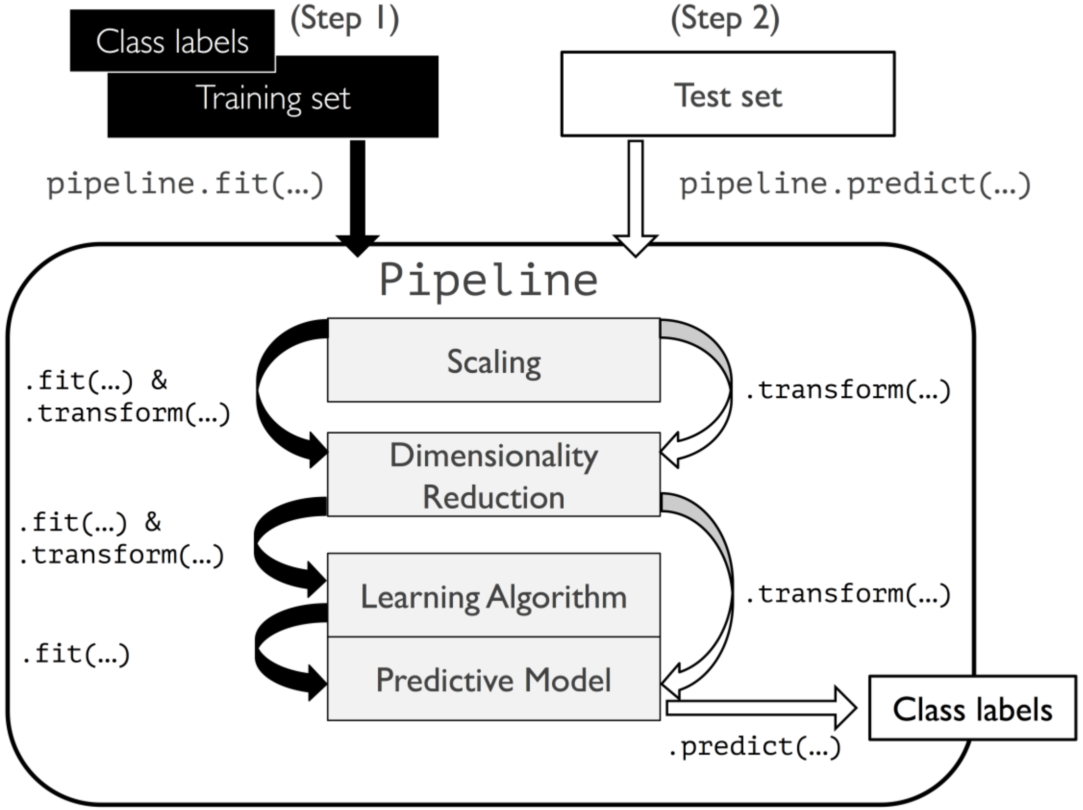
1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

今天課堂概要

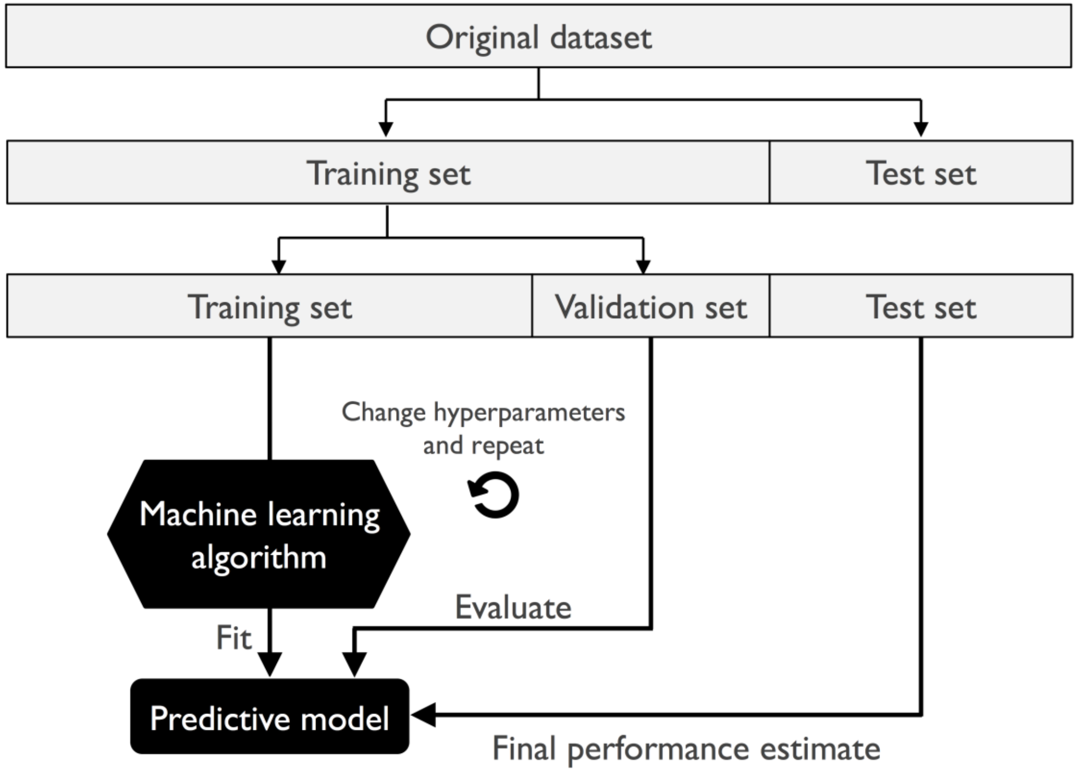
Model Evaluation

1. Pipeline & Validation (Holdout & k-fold)
2. Over- and underfitting addressed with validation curves
3. Evaluation matrix
4. Class imbalance

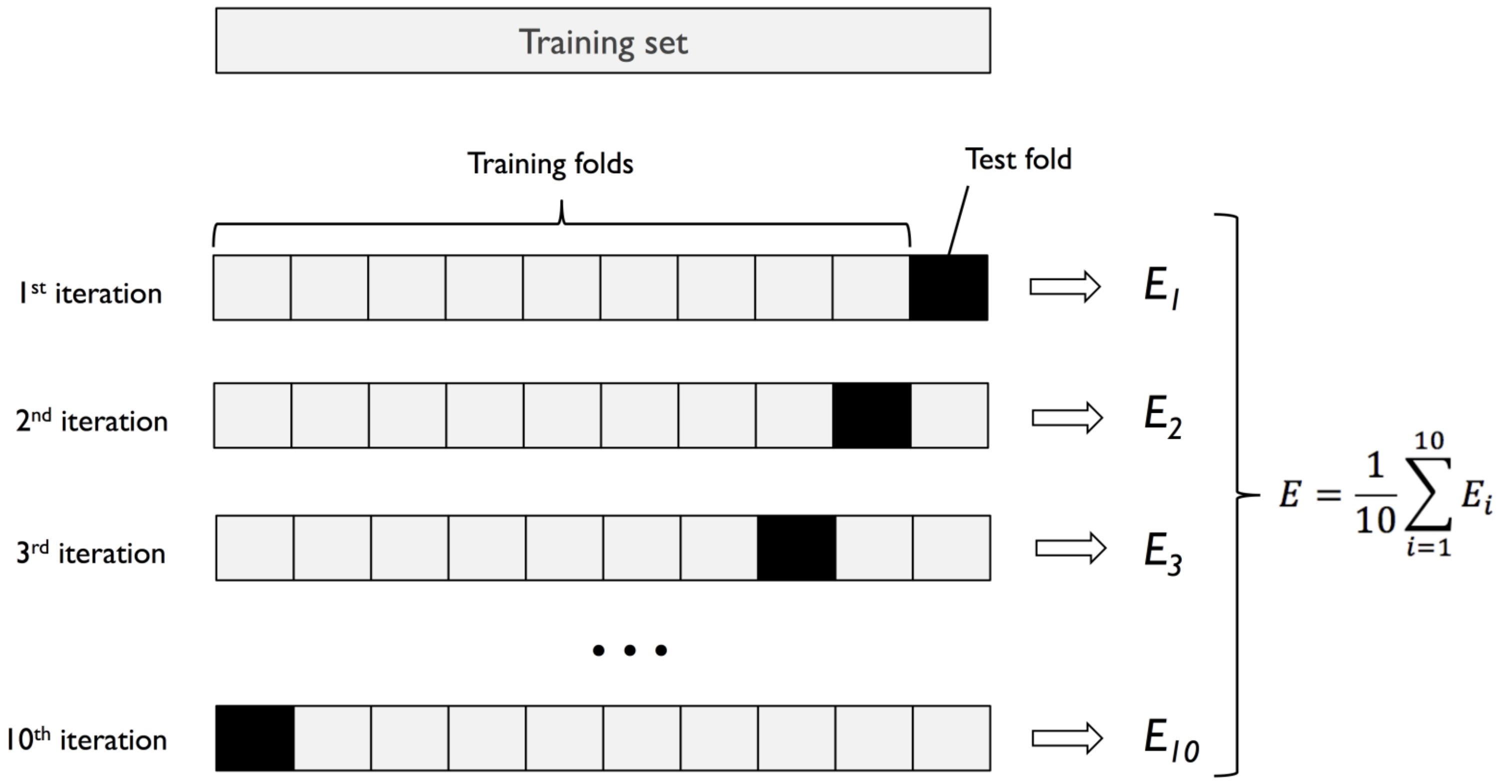
PIPELINE


Class labels
(Step 1)
Training set
(Step 2)
Test set
pipeline.fit(...)
pipeline.predict(...)
Pipeline
Scaling
.fit(...) &
.transform(...)
.transform(...)
Dimensionality
Reduction
.fit(...) &
.transform(...)
.transform(...)
Learning Algorithm
.fit(...)
Predictive Model
Class labels
.predict(...)

HOLDOUT VALIDATION



K-FOLD VALIDATION

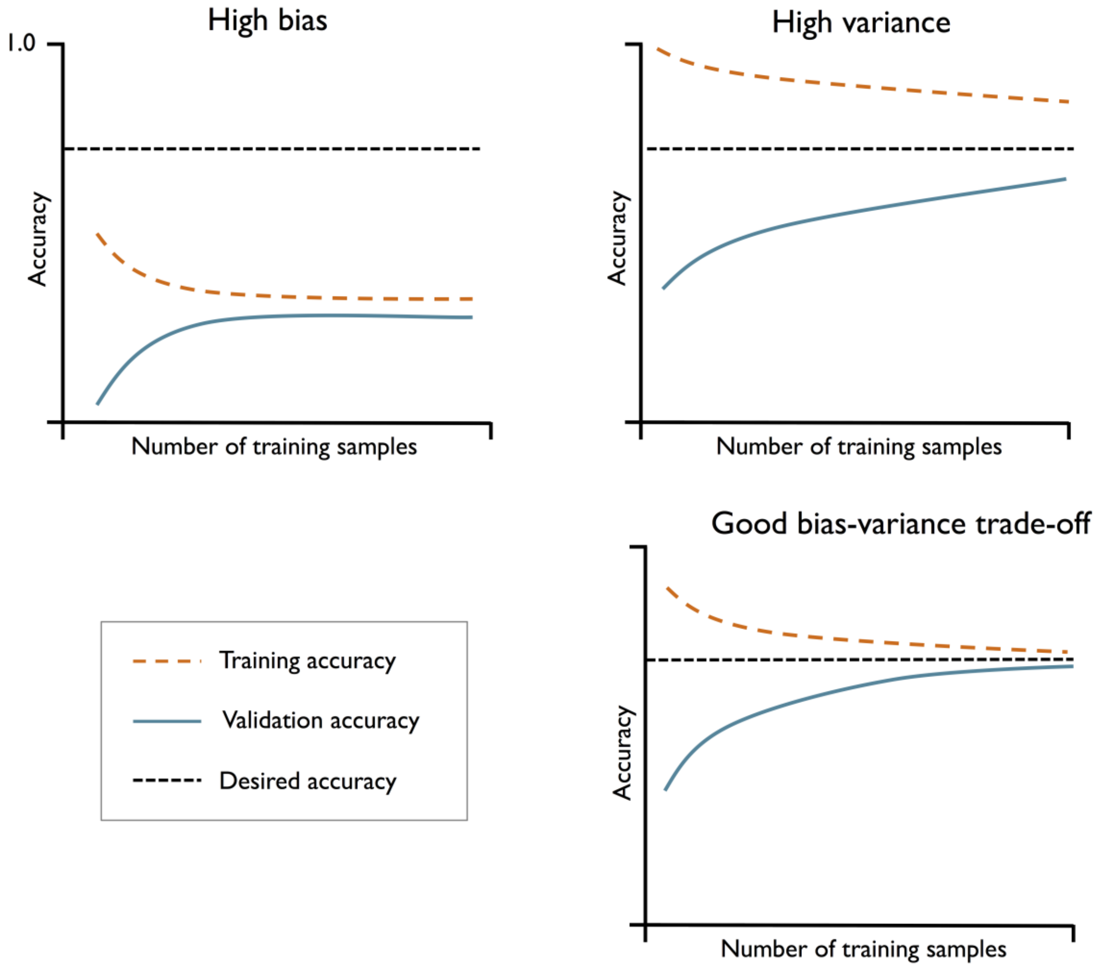


Intend to avoid the possible bias introduced by relying on any one particular division into test and train components, to partition the original set in several different ways and to compute an average score over the different partitions.

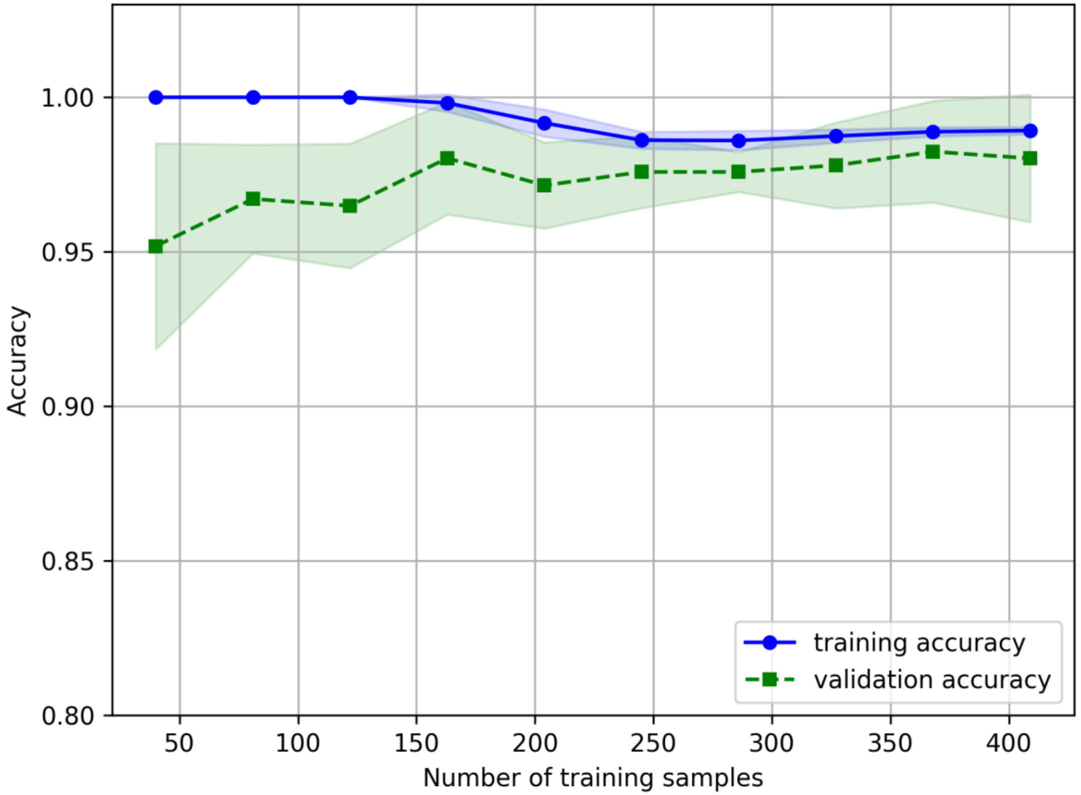
DIAGNOSING BIAS AND VARIANCE

High bias – Too restricted model

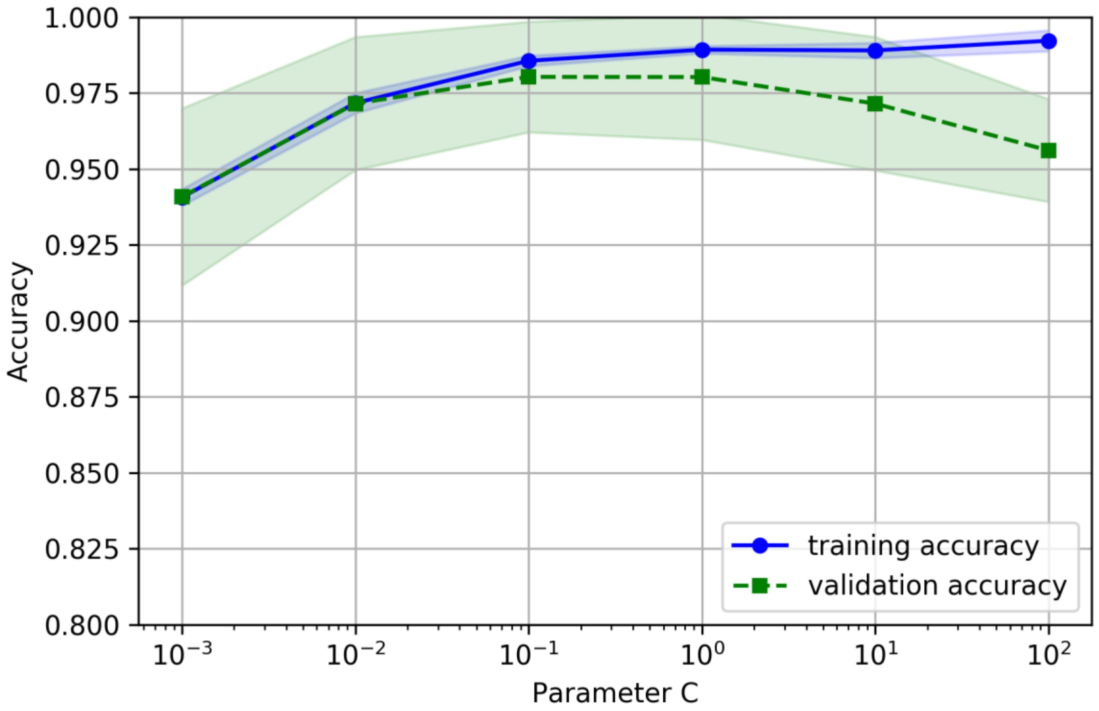
High variance – Too complicated model



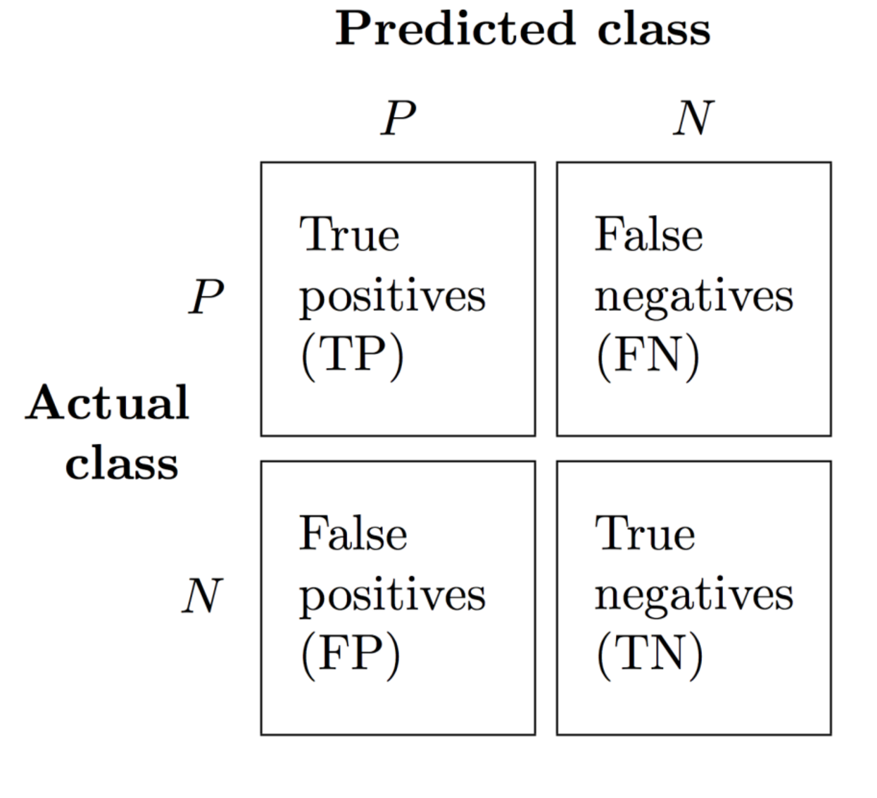
DIAGNOSING BIAS AND VARIANCE



OVER- AND UNDERFITTING ADDRESSED WITH VALIDATION CURVES



EVALUATION MATRIX



		Predicted class	
		P	N
Actual class	P	True positives (TP)	False negatives (FN)
	N	False positives (FP)	True negatives (TN)

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

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

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

$$\text{F1} = 2 \times \frac{PRE \times RECALL}{PRE+RECALL} = 2 \times \frac{1}{\frac{1}{Pre}+\frac{1}{Recall}}$$

"MISLEADING" ACCURACY

Suppose that there is a data set which has 1,000 data points (970 with false label and 30 with true label).

	Predict true	Predict false
Actual true	20	10
Actual false	70	900

Accuracy = (20+900)/(20+70+10+900) = 0.92

	Predict true	Predict false
Actual true	1	29
Actual false	0	939

Accuracy = 940/970= 0.96

EVALUATION MEASURES

Case 1:

Precision = 20/(20+70)=0.222

Recall = 20/(20+10)=0.667

F1 = 0.333

Case 2:

Precision = 0/(30)

Recall = 1/(1+30)=0.032

F1 = 0

IMBALANCED CLASS

Re-sampling

Class weighting

SMOTE

今天課堂概要

Model Evaluation

1. Pipeline & Validation (Holdout & k-fold)
2. Over- and underfitting addressed with validation curves
3. Evaluation matrix
4. Class imbalance

下一課…

1. Summarize the topics covered
2. Introduction of deep learning and demonstration of selected application (e.g. Text summarization or computer vision)