

應用機器學習

Brian Chan 陳醒凡

課程目標

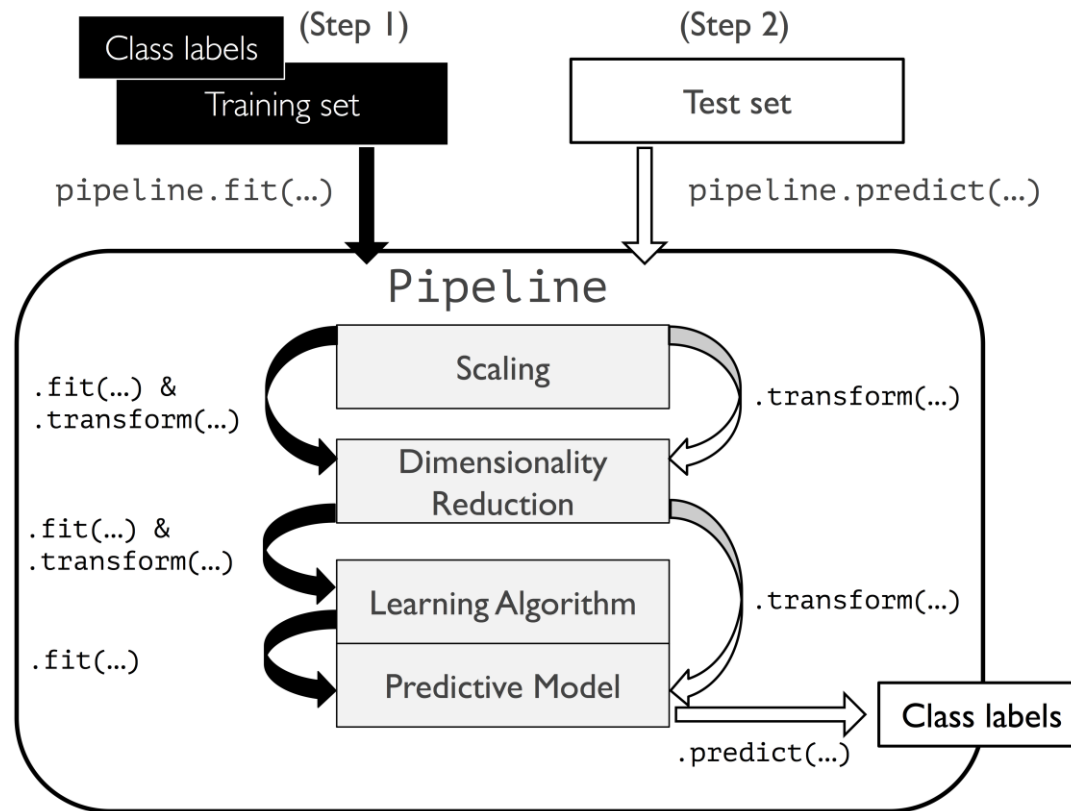
1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

今天課堂 概要

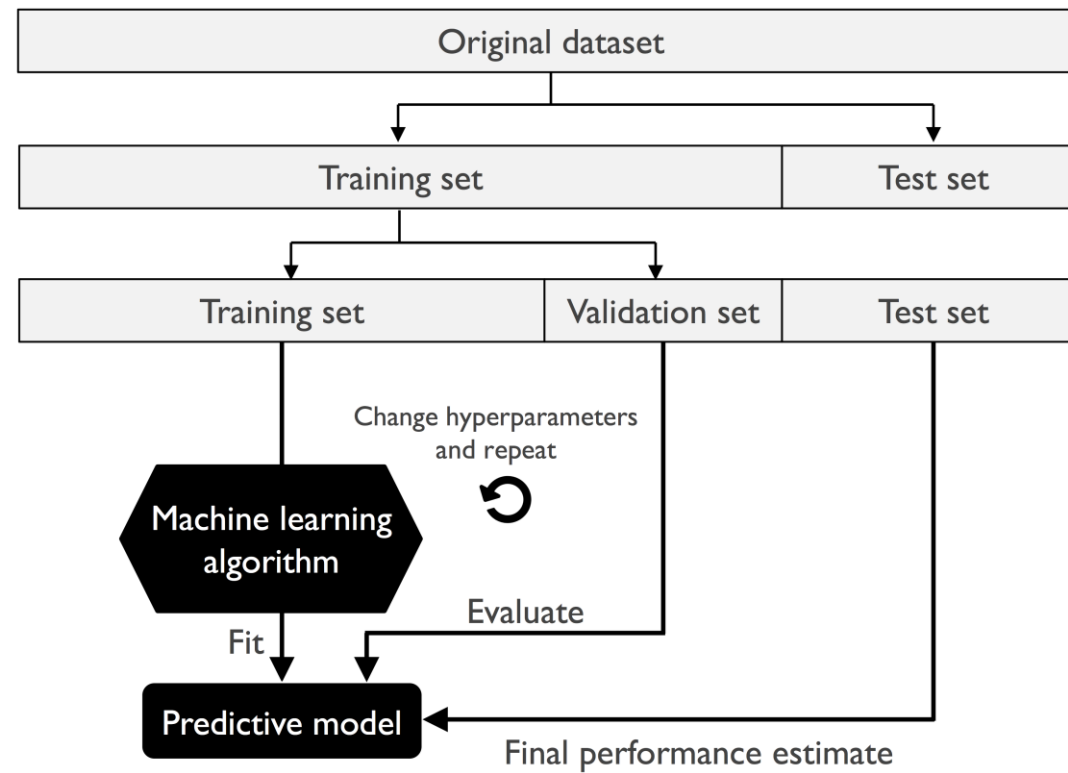
Model Evaluation

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

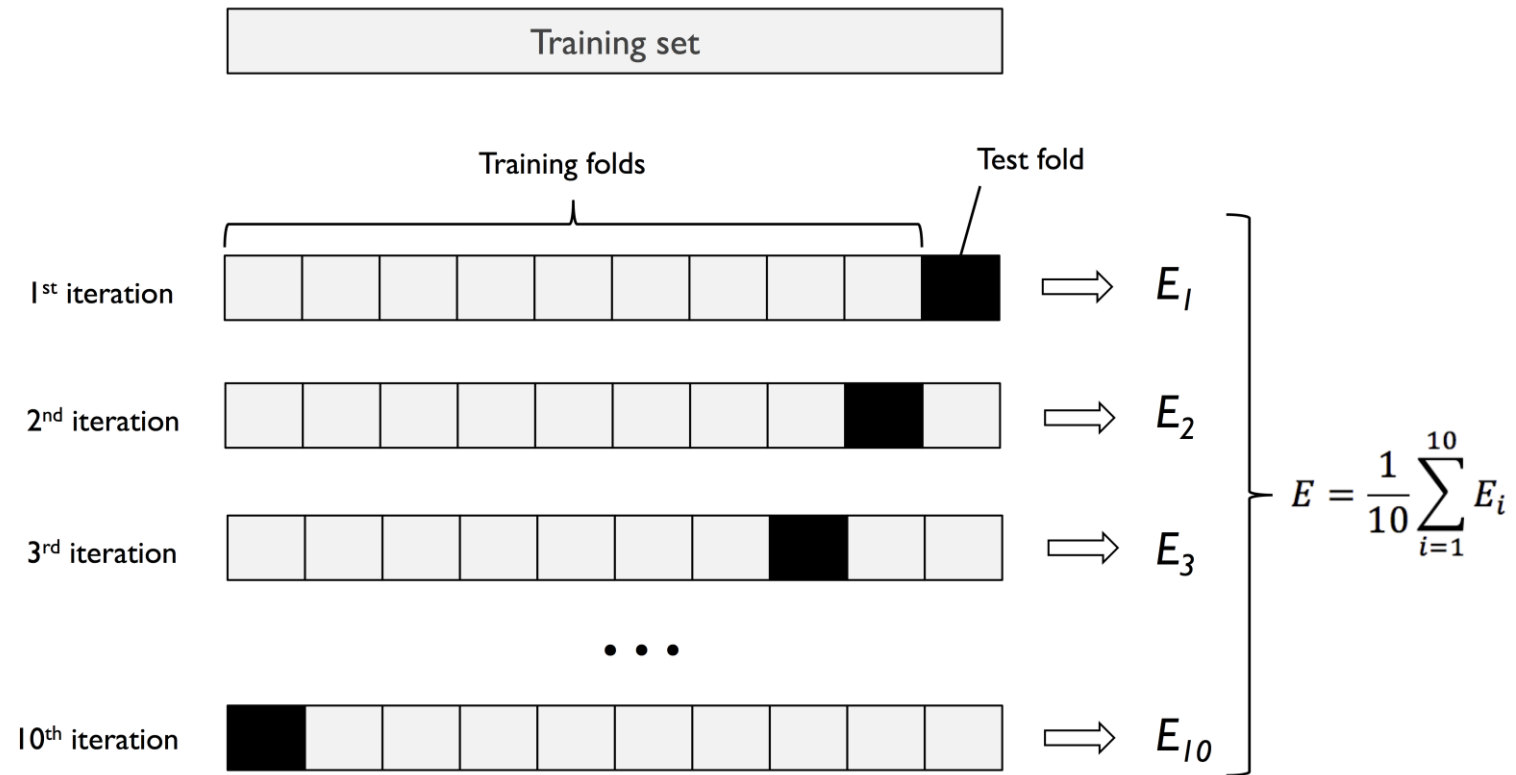
PIPELINE



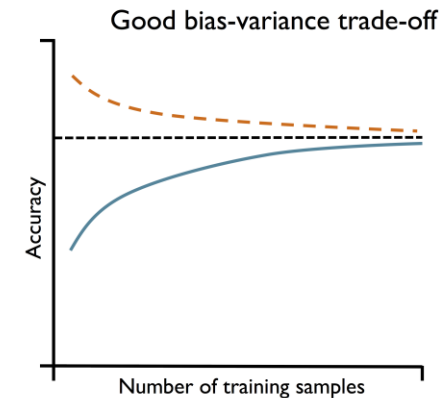
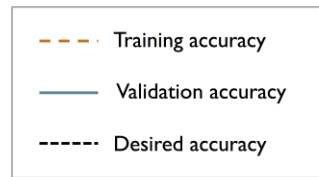
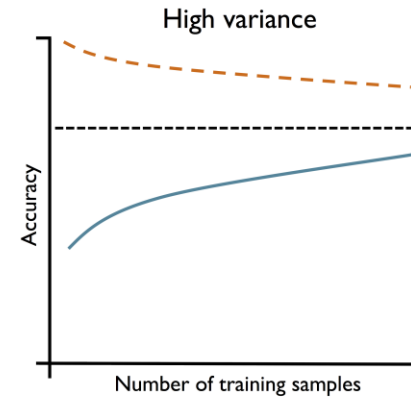
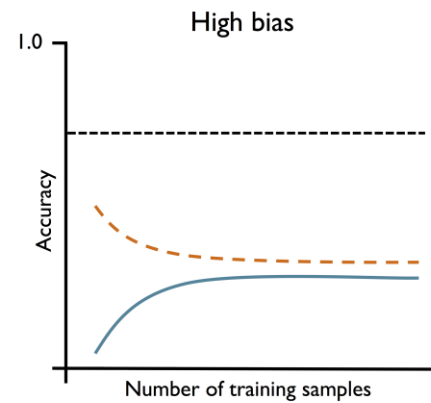
HOLDOUT VALIDATION



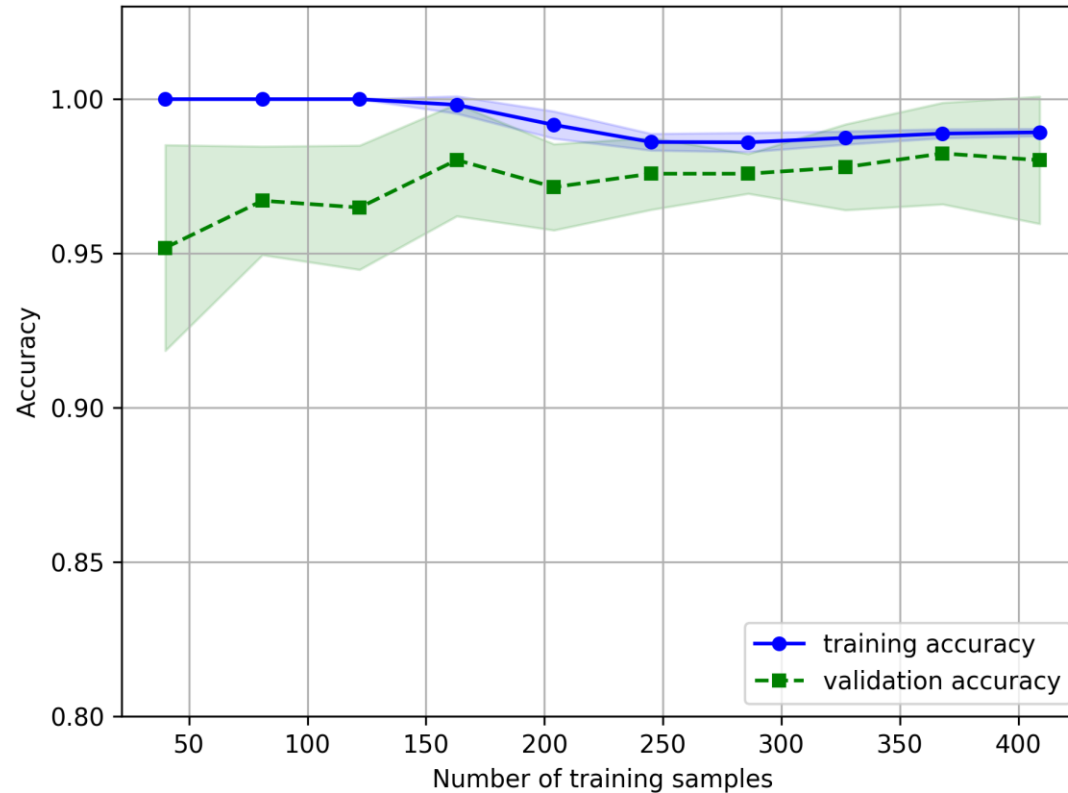
K-FOLD VALIDATION



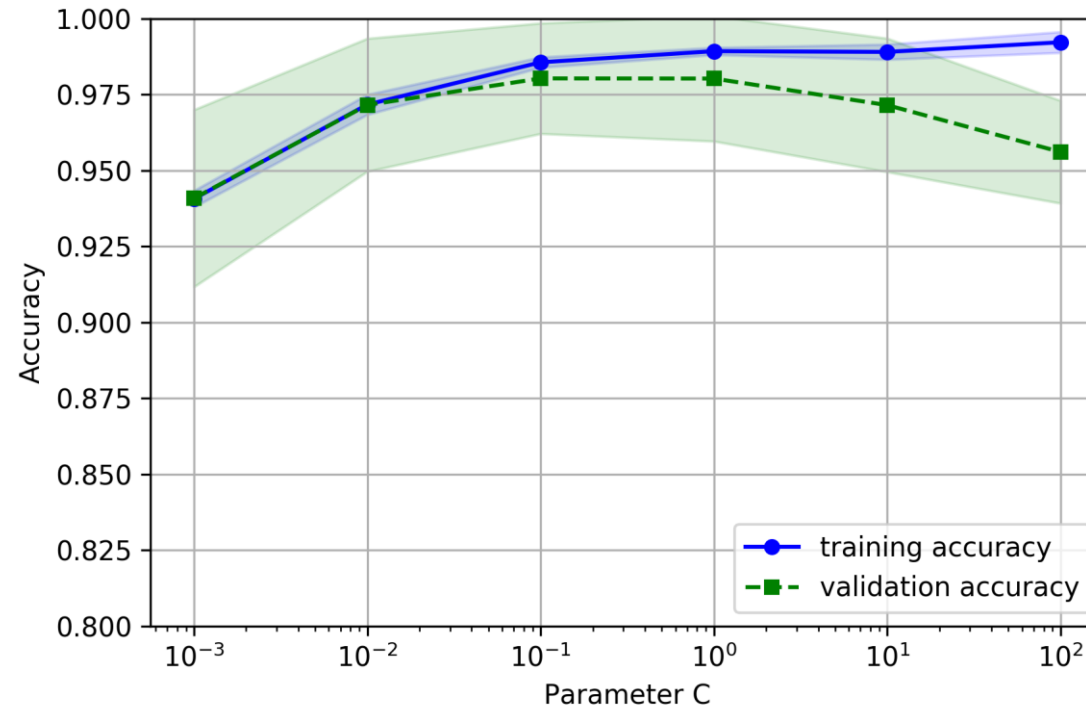
DIAGNOSING BIAS AND VARIANCE



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)

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

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

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

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

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

$$\text{Accuracy} = (20+900)/(20+70+10+900) = 0.92$$

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

$$\text{Accuracy} = 940/970 = 0.96$$

Case 1:

$$\text{Precision} = 20/(20+70)=0.222$$

$$\text{Recall} = 20/(20+10)=0.667$$

$$\text{F1} = 0.333$$

Case 2:

$$\text{Precision} = 0/(30)$$

$$\text{Recall} = 1/(1+30)=0.032$$

$$\text{F1} = 0$$

UNBALANCED DATA

Re-sampling

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)