Programming for Social Scientists: Text Mining VI

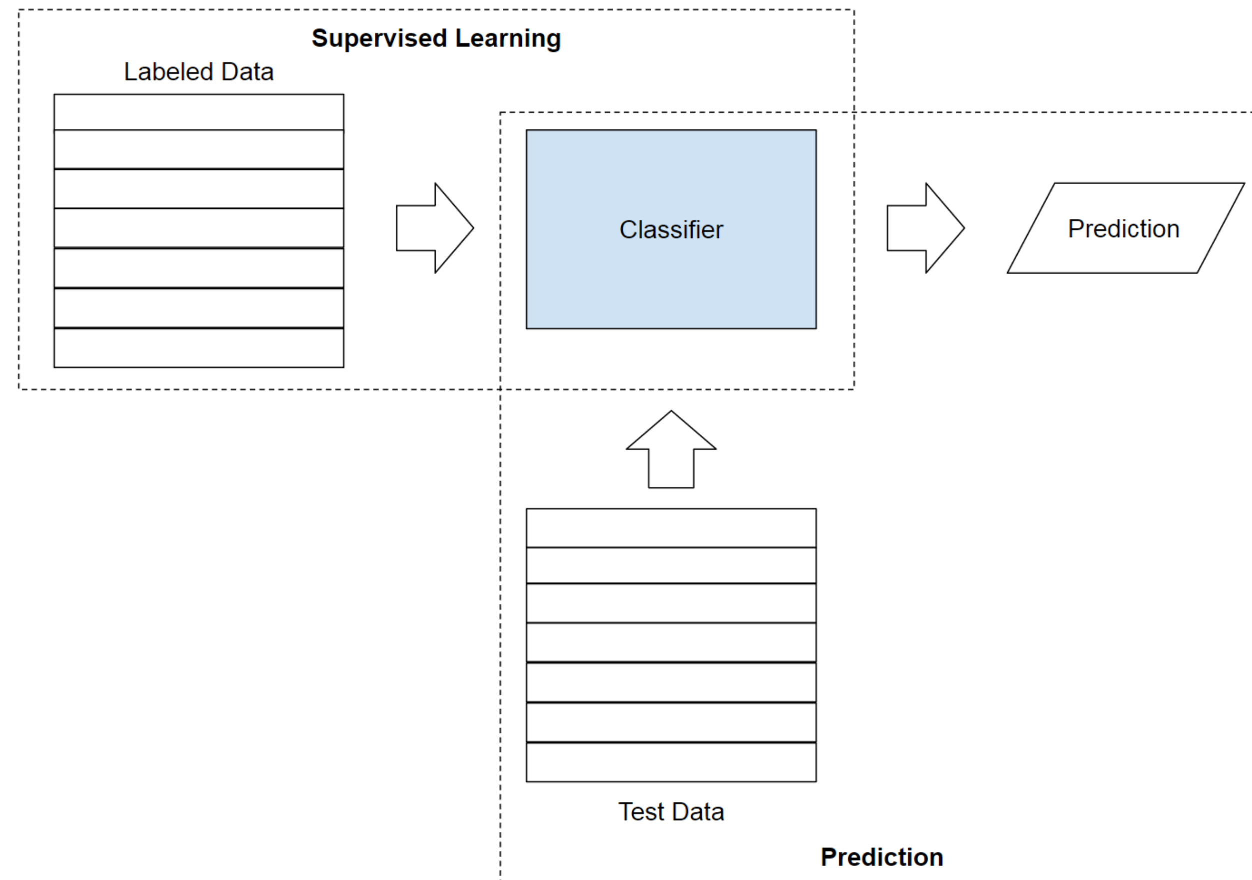
黃瀚萱

Agenda

- Classification
- Evaluation
- Applications

Supervised Learning

- To obtain a model *f*() that outputs *y* given an input *x*.
    - $f(x) = y$
- Learning from labeled data
    - To capture the relation between (x, y) with a large amount of (x, y) pairs.



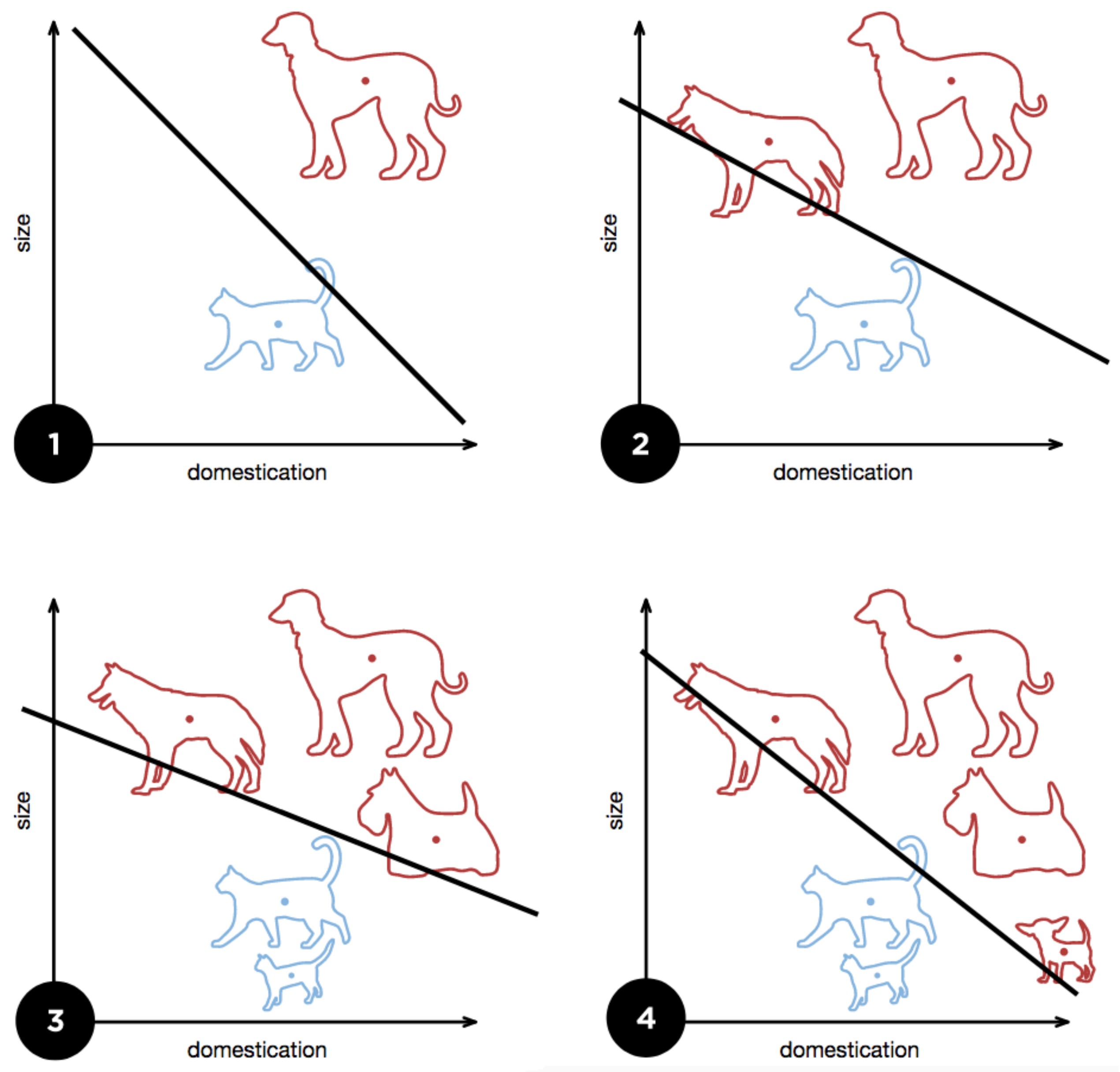
Perceptron

Features ($\mathbf{x}$):
x_1: size
x_2: domestication

Labels (y):
0: Dog
1: Cat

$$f(x) = \begin{cases} 1 & \text{if } w_1 x_1 + w_2 x_2 + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

Parameters to estimate: w_1, w_2, and b



Training the Perceptron model

$D = \{(x_{1,1}, x_{1,2}, x_{1,3}, ..., x_{1,n}, y_1), (x_{2,1}, x_{2,2}, x_{2,3}, ..., x_{2,n}, y_2),$
$..., (x_{m,1}, x_{m,2}, x_{m,3}, ..., x_{m,n}, y_m)\}$

For convenient, set $x_{i,0} = 1$ for all input $1 \leq i \leq m$,
and use w_0 as the bias instead of b.

Initialization
$w_j = 0$ for all weights $0 \leq j \leq n$.

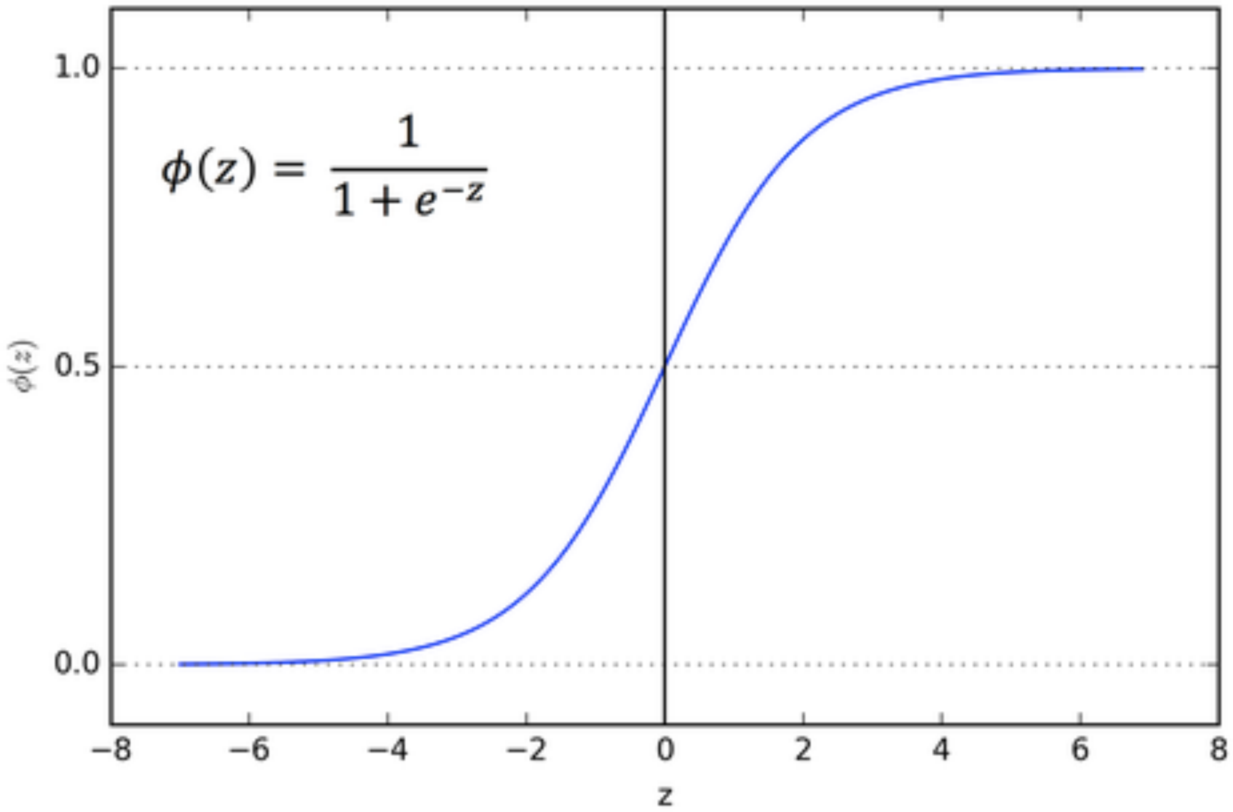
Training
Do until convergence
    For each instance $(\mathbf{x}_i, y_i)$ in D
        Evaluate $f(\mathbf{x}_i)$
            $y'_i \leftarrow w_0 x_{i,0} + w_1 x_{i,1} + x_2 x_{i,2} + ... + w_n x_{i,n}$
        Update weights
            $w'_j \leftarrow w_j + \alpha(y_i - y'_i)x_{i,j}$ for all features $0 \leq j \leq n$.

Logistic Regression

- Estimate the probability of a binary outcome (y) based on input features x.

- Map the output of perceptron f(x) to the range (0, 1) with the logistic (sigmoid) function.

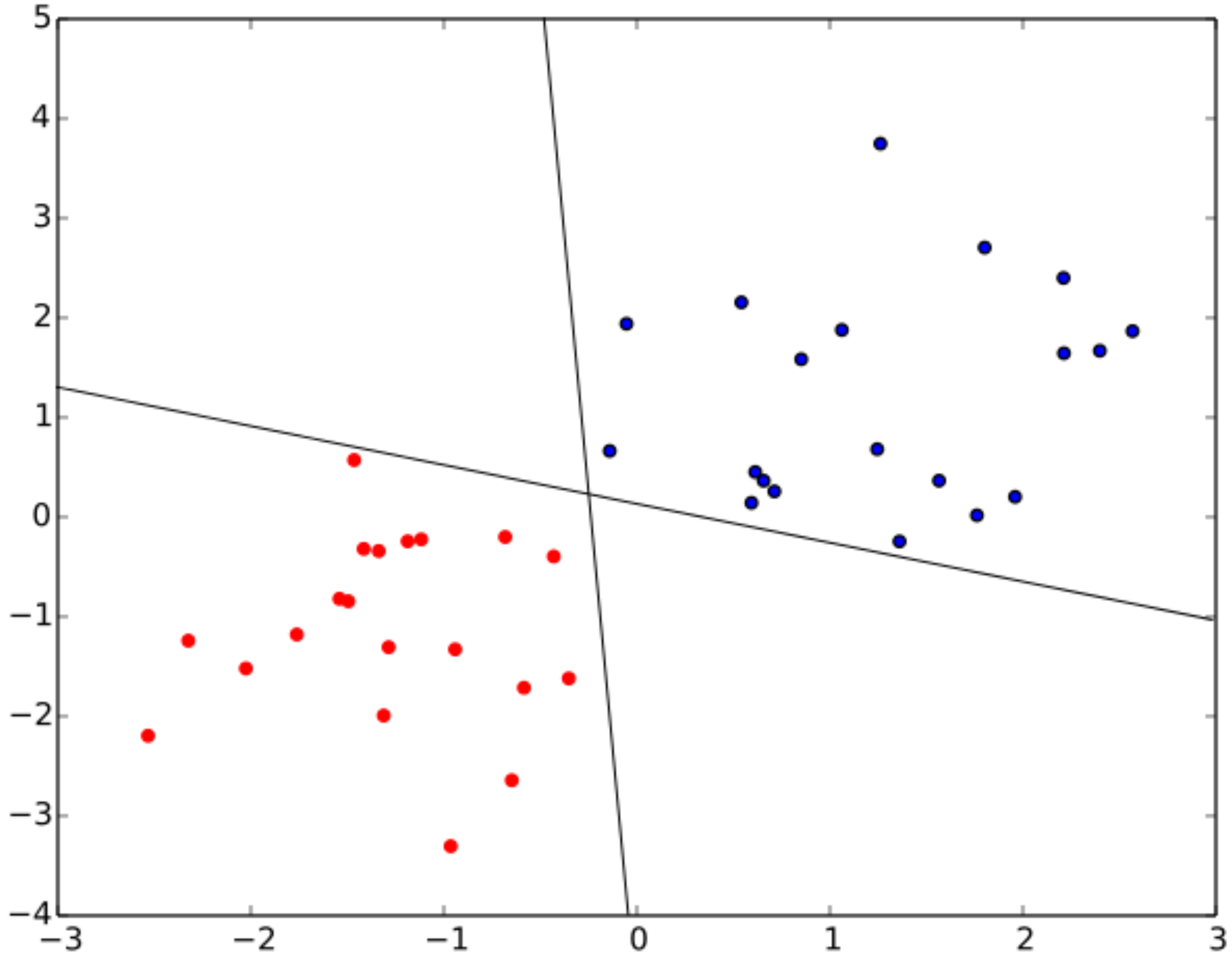
$$p(f(x) = 1)$$
$$= \frac{1}{1 + e^{-f(\mathbf{x})}}$$
$$= \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n)}}$$



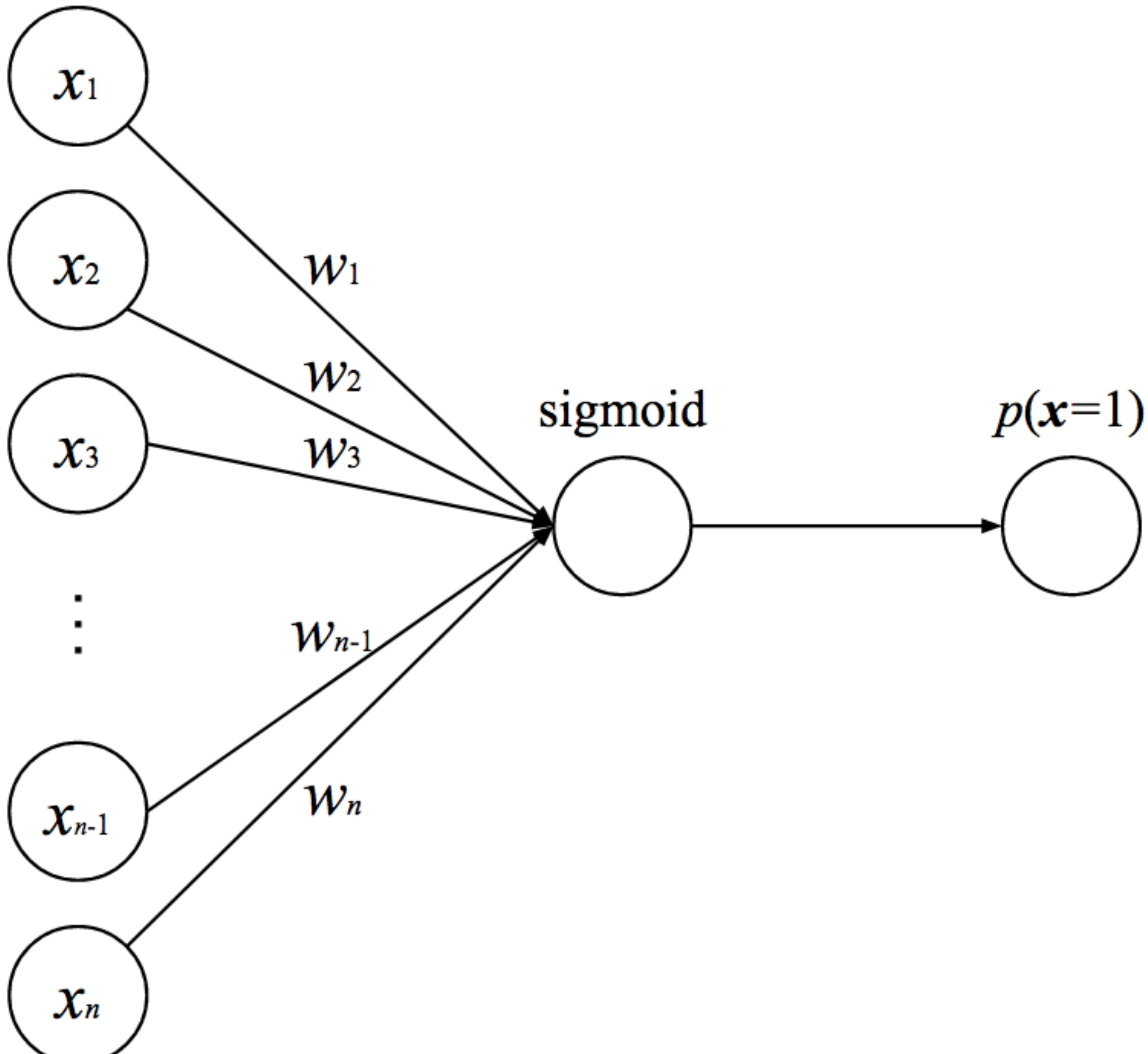
Maximum Margin

- Perceptron cannot choose the best splitting in the training space.

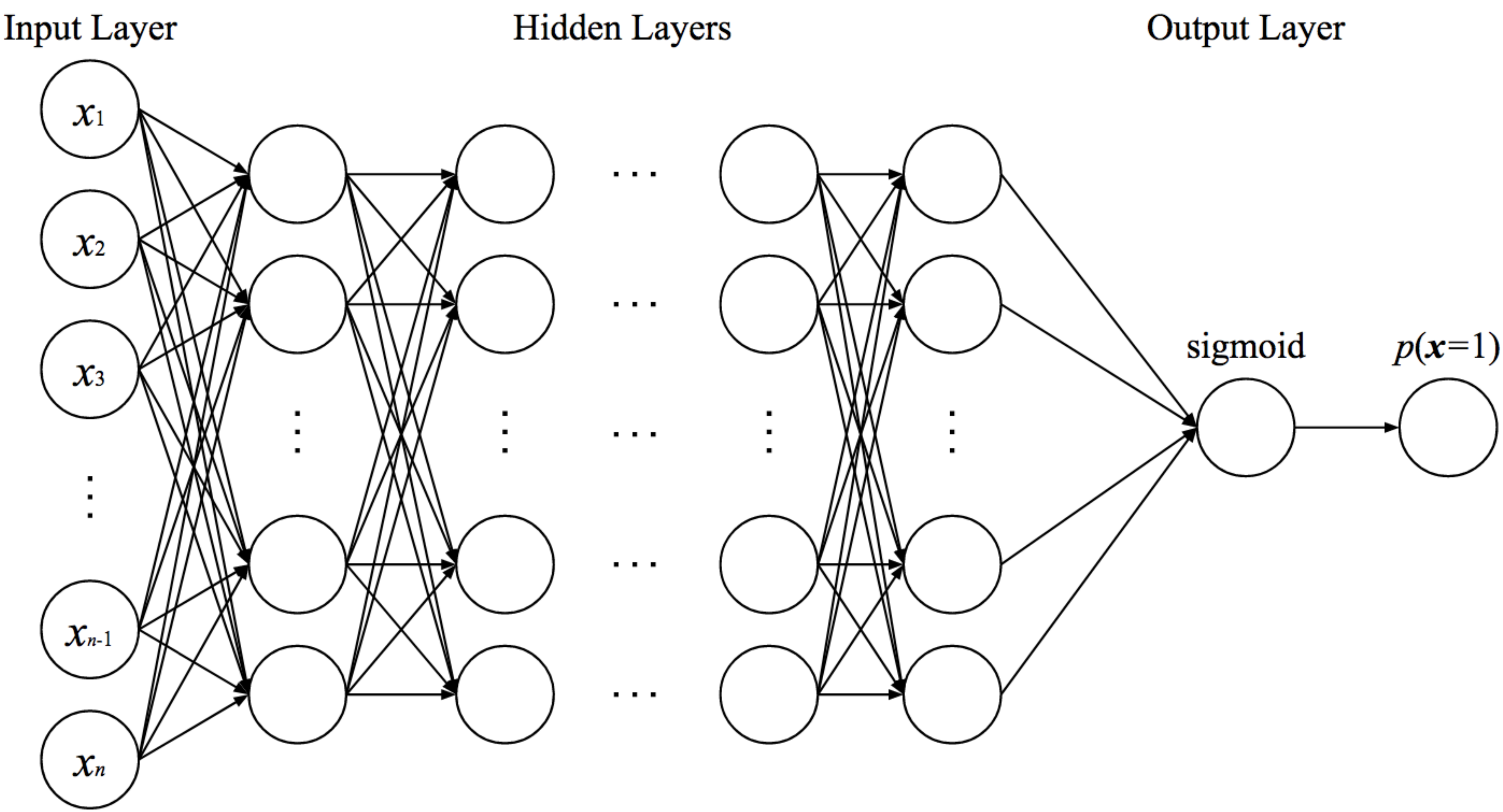
- Maximum margin models like SVM are designed to solve this problem.



Single Layer Neural Network

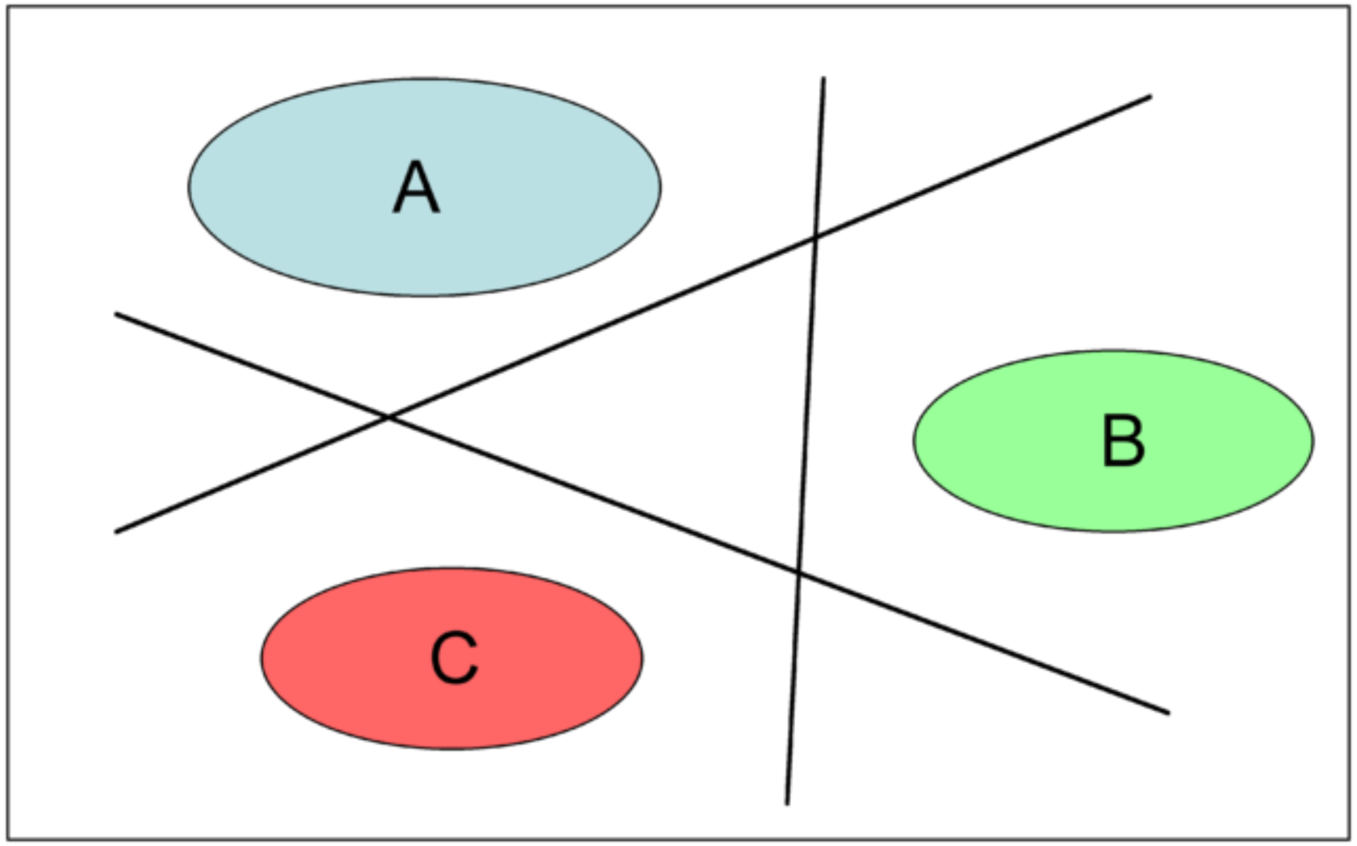


Deep Neural Network



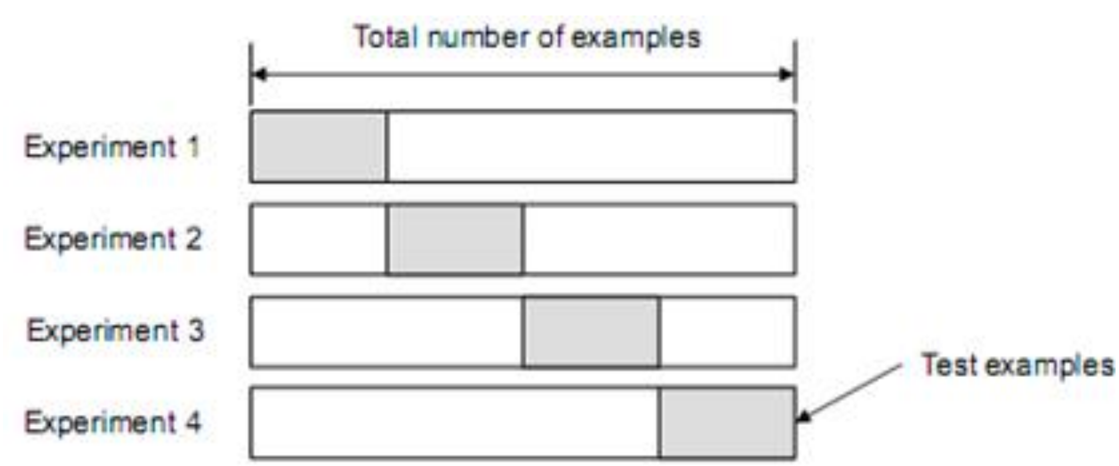
Multiclass Classification

- Multiclass: Cat vs Dog vs Pig
- One vs Rest method
  - Cat or Not-Cat
  - Dog or Not-Dog
  - Pig or Not-Pig



Performance Evaluation

- Holdout
  - Keeping a set of labeled data from the training set to simulate the new instances in the real application scenario.
- Cross-validation
  - Splitting labeled data into *k* folds
  - Each *i*-fold is used as test set, and the instances in other folds are used as training data.



Metrics

- 假設有一個在海關偵測入境旅客有無發燒的模型
  - True-Positive: 模型判定為發燒，旅客也確實發燒
  - True-Negative: 模型判定為無發燒，旅客也確實無發燒
  - False-Positive: 模型判定為發燒，但旅客其實沒有發燒 (干擾旅客)
  - False-Negative: 模型判定為無發燒，但旅客其實有發燒 (公衛問題)

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F - score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

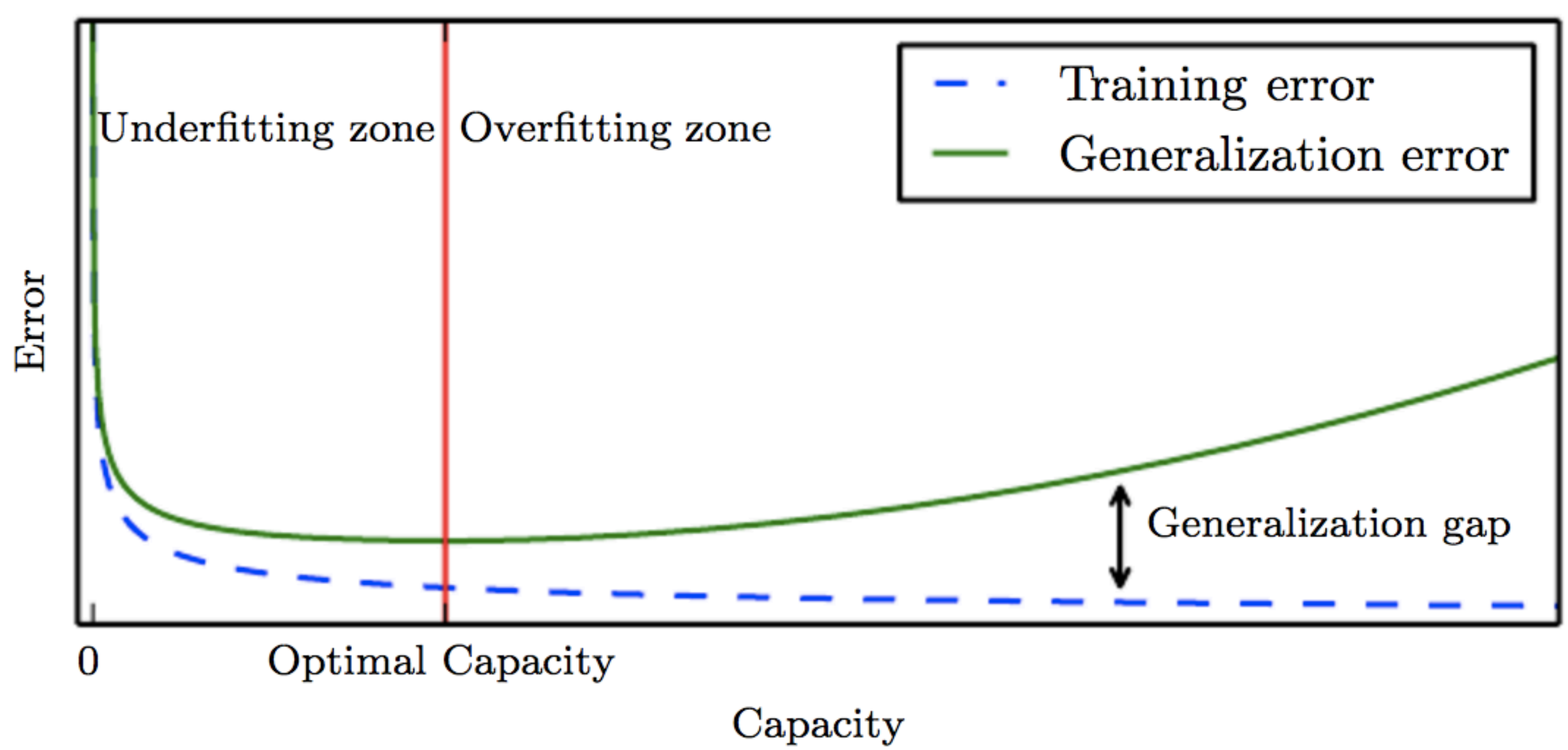
		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Recall vs Precision

- Accuracy代表整體分類正確率，但易受資料不平衡影響
  - 假設95%的旅客都沒有發燒，模型一律預測無發燒就有95%的accuracy。
- Precision反應出模型認為有發燒的旅客，究竟多少真的發燒。
  - 從嚴判定，要40度以上才認為發燒 Precision 會很高，但Recall會偏低
- Recall反應出模型可以找出多少真正發燒的旅客
  - 一律判定為發燒，Recall=100%，但 Precision只有5%
- F-Score 是Precision與Recall的調和平均數，可以同時考慮兩者的平衡。

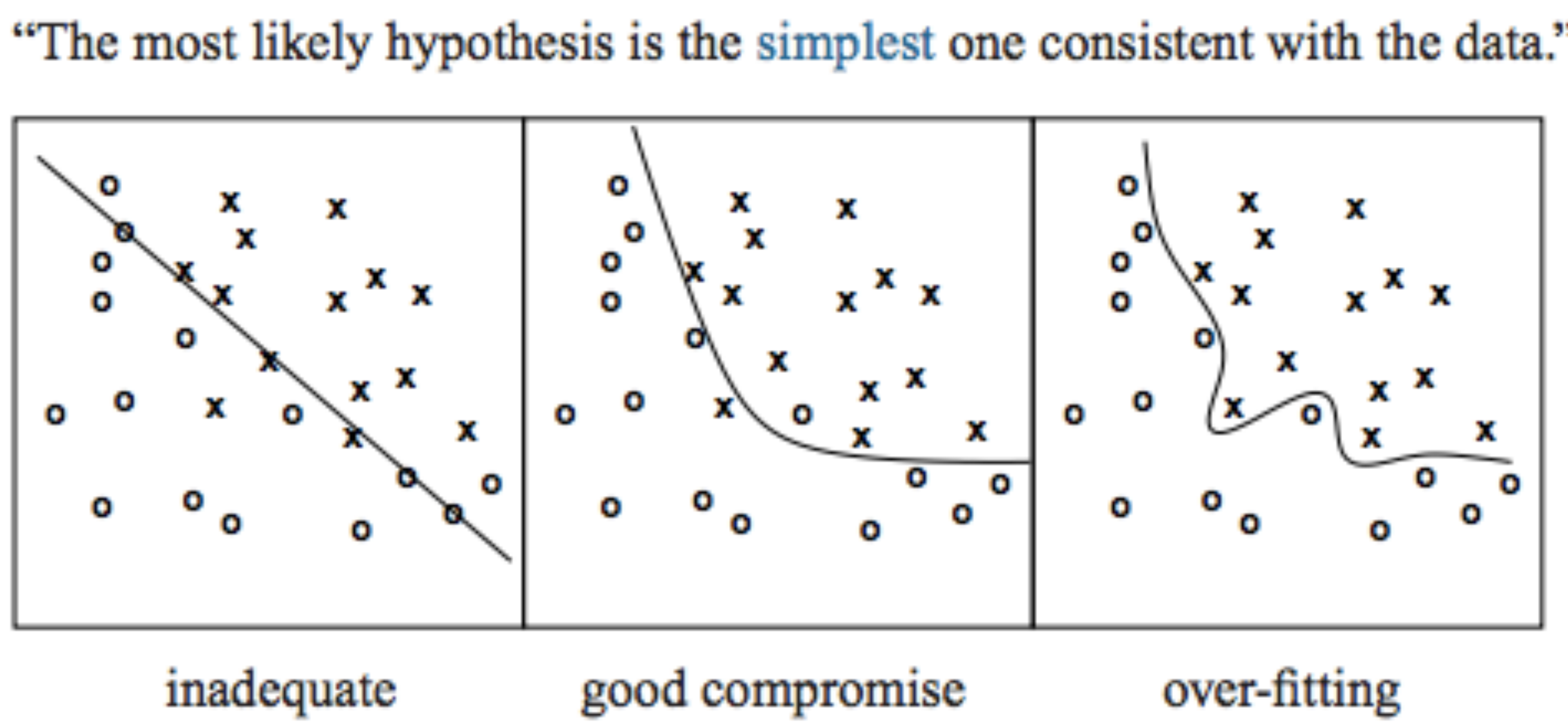
Training Performance vs Validation Performance

- Training performance
  - Used to measure model's ability to fit the data.
- Testing (validation) performance
  - Used to measure the performance in real applications.



Overfitting

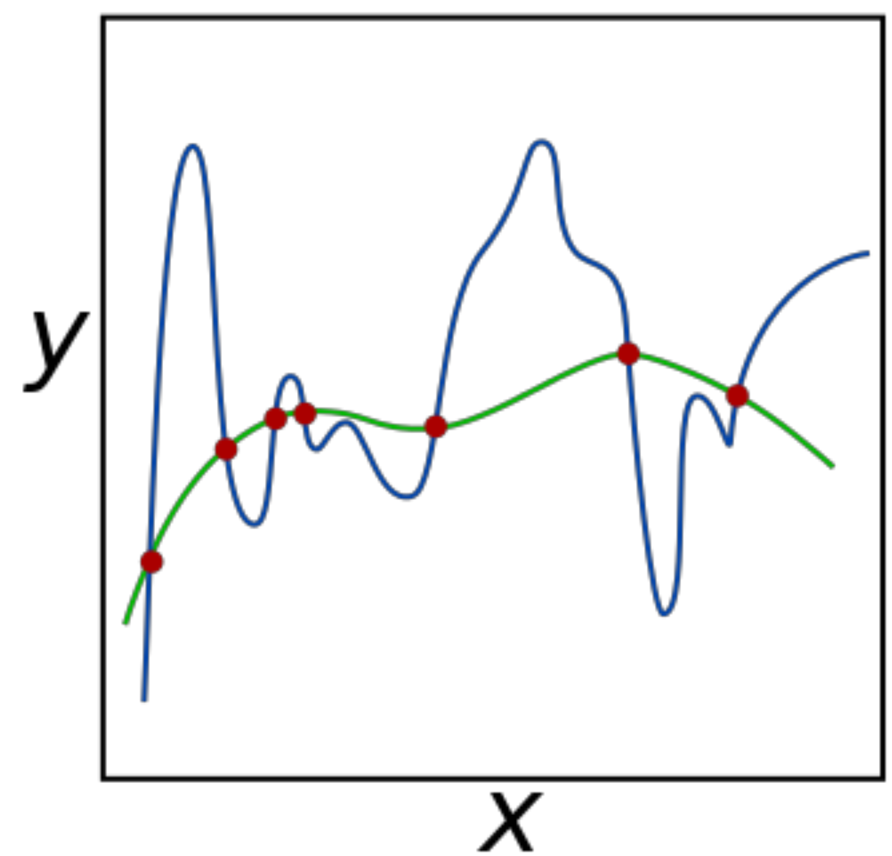
- Training performance >> testing performance



Data Sparseness

- Reasons of overfitting
  - Too specific features
  - Too complex model
- Solutions
  - More training data
  - Reduce the complexity of the model
  - Feature selection
  - Regularization

$$
\begin{aligned}
D = \{ \\
&(x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n}, y_1), \\
&(x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n}, y_2), \\
&\ldots, \\
&(x_{m,1}, x_{m,2}, x_{m,3}, \ldots, x_{m,n}, y_m)\} \\
, \text{ where } n \gg m
\end{aligned}
$$

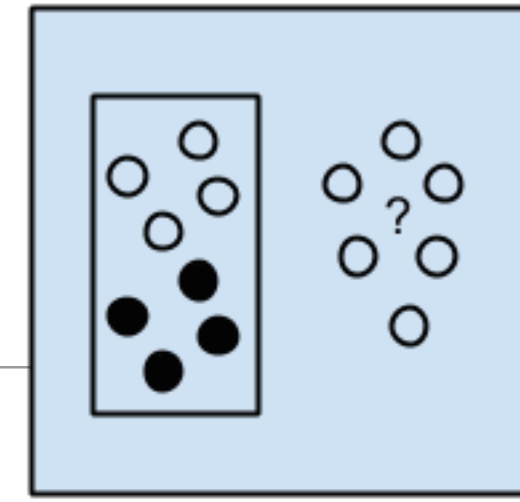


Types of Machine Learning

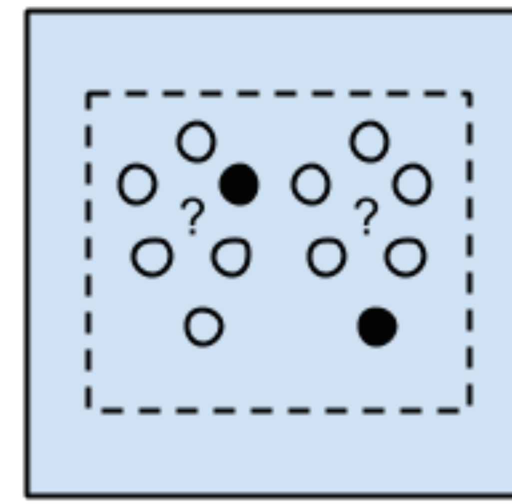
- Types of output
  - Nominal (Categorical): Classification (Cat vs Dog)
  - Continuous: Regression (Stock price)
- Availability of labeled data
  - Supervised learning: With labeled data
  - Unsupervised learning: Without labeled data
  - Semi-supervised: Some of data are labeled

Supervised vs Unsupervised

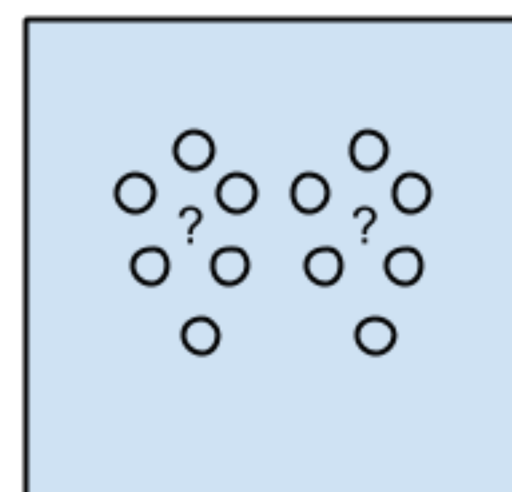
- Supervised
  - Train the model with labeled instances.
- Unsupervised
  - Clustering
- Semi-supervised
  - Improve the supervised learning with unlabeled data.



Supervised Learning Algorithms



Semi-supervised Learning Algorithms



Unsupervised Learning Algorithms

Applications

- Gaming（Google AlphaGo）
- Question answering（IBM Watson）
- Auto driving (Google/Uber/Apple/Tesla)
- Machine translation (Google Translate)
- Intelligent assistance (Apple Siri/Amazon Echo)
- Advertising (Google/Facebook/Criteo/Tagtoo)

Deceptive Information Detection

- Spam (Gmail)
- Phishing
- Paid writer
- Exaggerated advertisement
- Water army
- Clickbait
- Robot

Assignment

- Do text classification with both kinds of features.
  - Bag-of-words
  - Word embeddings
- Define your representation method for a better performance.

One-hot Features													Word embeddings						
0	1	0	1	0	0	1	0	0	1	0	0	0	.	.	.	.	.	.	.