Classification

Data Mining:
Data Mining Methods
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

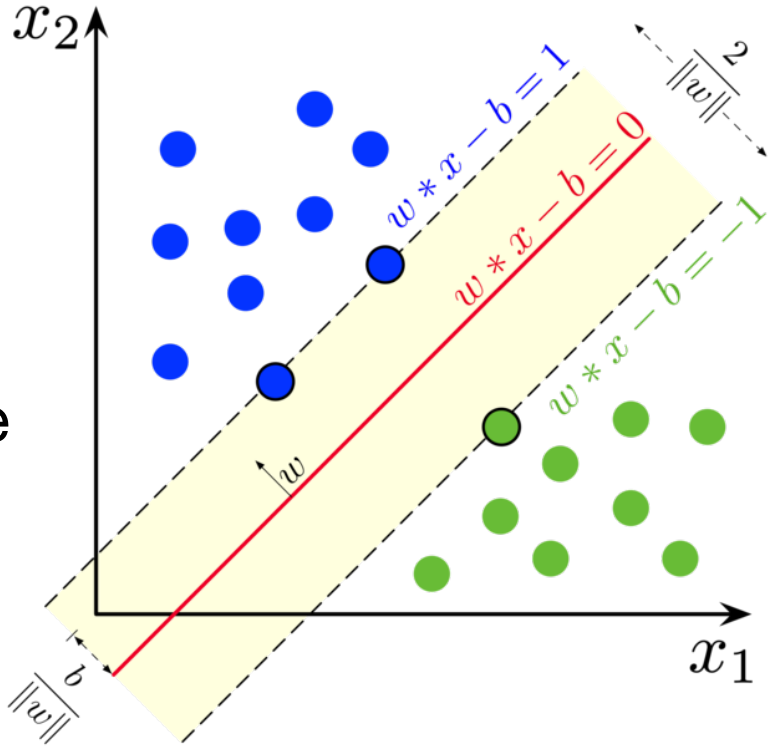
Learning objective: Apply techniques for classification and explain how they work. Evaluate and compare methods.

Classification

- ➢ Supervised learning
  - Training set with predefined class labels
- ➢ Decision tree induction
  - Top-down, recursive, attribute selection & split
- ➢ Bayesian classification
  - Probability, naïve assumption, belief network

Support Vector Machines (SVM)

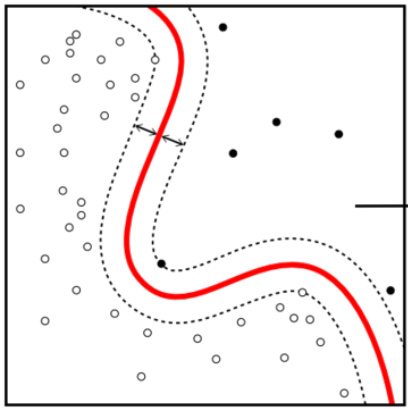
- ➢ Objects w/ class label
- $(X_1, y_1), \ldots, (X_n, y_n)$
- ➢ Separating hyperplane
- Maximum margin
- Maximum margin hyperplane
- Support vectors



SVM: Linear Separability

➢ Linearly separable

- Original space

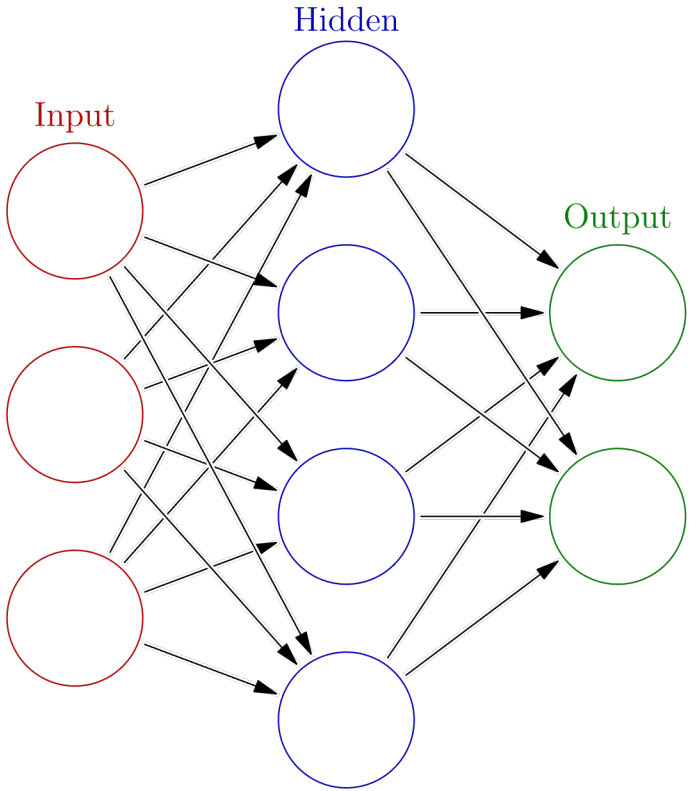


➢ Linearly inseparable

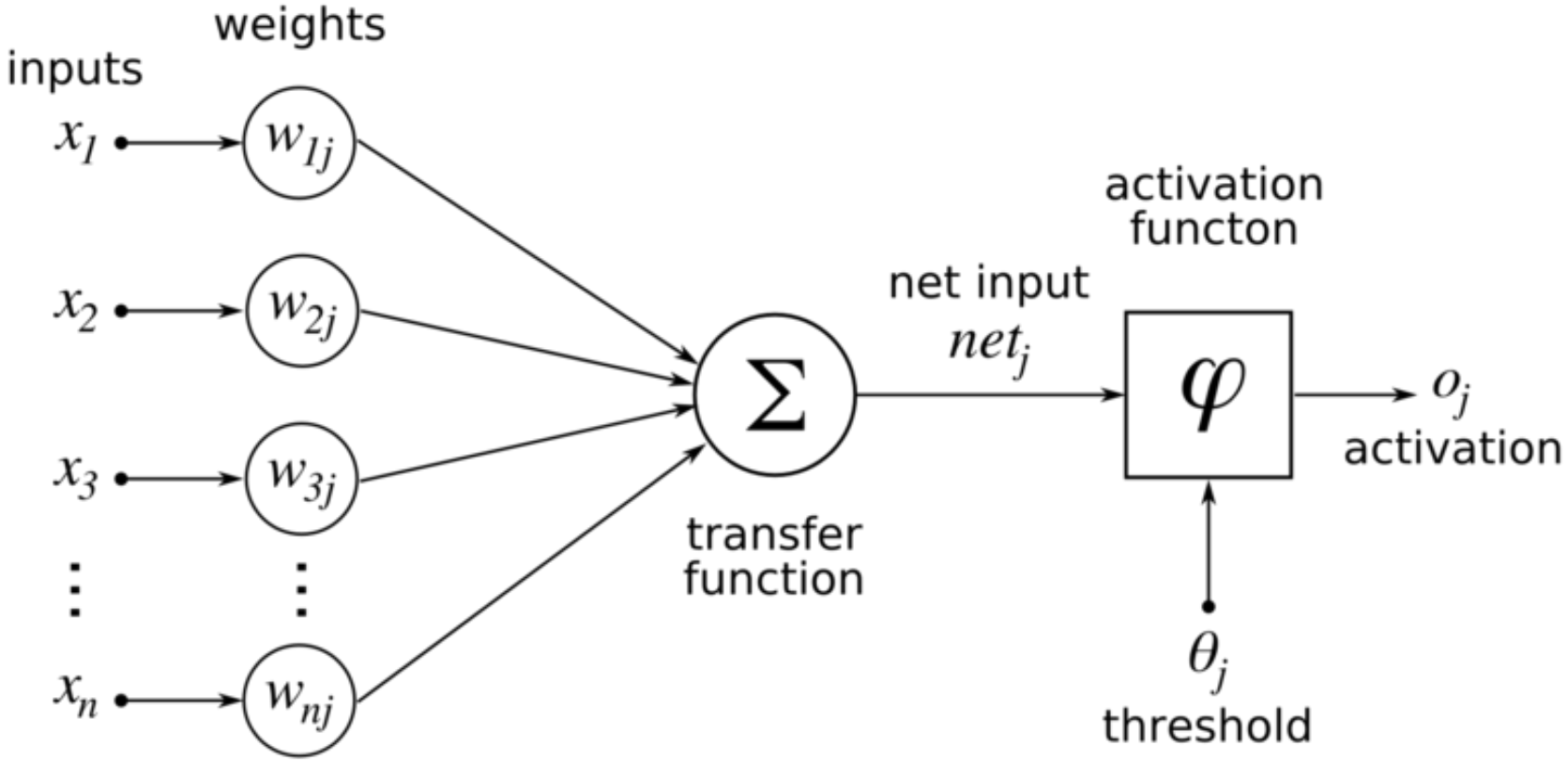
- => higher dimension space
- Dot product on transformed data is mathematically equivalent to applying a kernel function to original data

Neural Network

- ➢ Input, hidden, output layers
- • #layers, #units/layer
- ➢ Weighted connections
- • Initial weights, adjustments
- ➢ Feedforward
- • Observations => classification
- ➢ Backpropagation
- • Classification error => weight adjustment

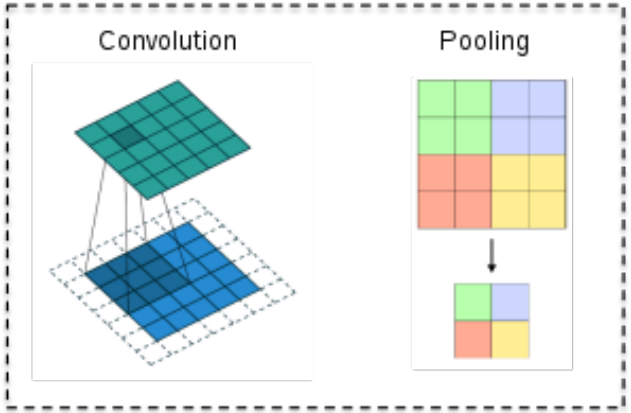


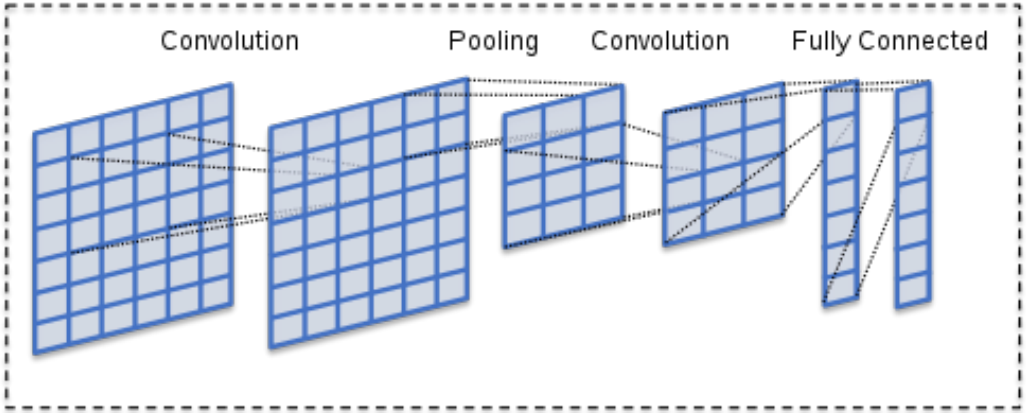
Neuron: Hidden/Output Layer Unit



Deep Neural Network

- Advances in data, computation, model
- E.g., convolutional neural network (CNN)
  - Activation function, regularization, attention, …





Classification Methods

- Decision tree induction: Efficient, easy to interpretate
- Bayesian classification
  - Efficient, explainable, incremental
- Support vector machines: Good/high performance
- Neural networks
  - Complex, high performance, poor interpretability

Ensemble

- ➢ Combined use of multiple models
  - Analogy: consulting multiple medical doctors
- ➢ Bagging: equal weights, majority voting
  - Training set: random sample with replacement
- ➢ Boosting: weighted votes
  - Adjust weights to focus more on misclassified cases

Model Evaluation

- ➢ Holdout, random sampling
  - Split into training set (for model construction) & test set
- ➢ (k-fold) Cross-validation
  - Split into k partitions, each 1 for testing & (k-1) for training
- ➢ Bootstrapping: (e.g., .632 bootstrapping)
  - Random sample with replacement

Classification Accuracy (1)

- Confusion matrix
- E.g., Fraud detection
- Sensitivity: t_pos / pos
- Specificity: t_neg / neg
- Precision: t_pos / (t_pos + f_pos)
- Accuracy:

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive	False Negative
	No	False Positive	True Negative

$$accuracy = sensitivity \frac{pos}{pos + neg} + specificity \frac{neg}{pos + neg}$$

Classification Accuracy (2)

➢ Costs and benefits of TP, TN, FP, FN

- E.g., fraud detection, medical diagnosis
- False positive: a normal case is flagged as fraud
- False negative: a fraud is misclassified as normal

➢ Multi-class classification

- Exact match: i.e., predicted class = actual class
- Some classes may be more similar (or different)

Prediction Error

- E.g., stock price, travel time
- Difference between predicted values and actual values

$$Absolute\ error : |y_i - y_i'|$$

$$Square\ error : (y_i - y_i')^2$$

$$Mean\ absolute\ error : \frac{\sum_{i=1}^{d} |y_i - y_i'|}{d}$$

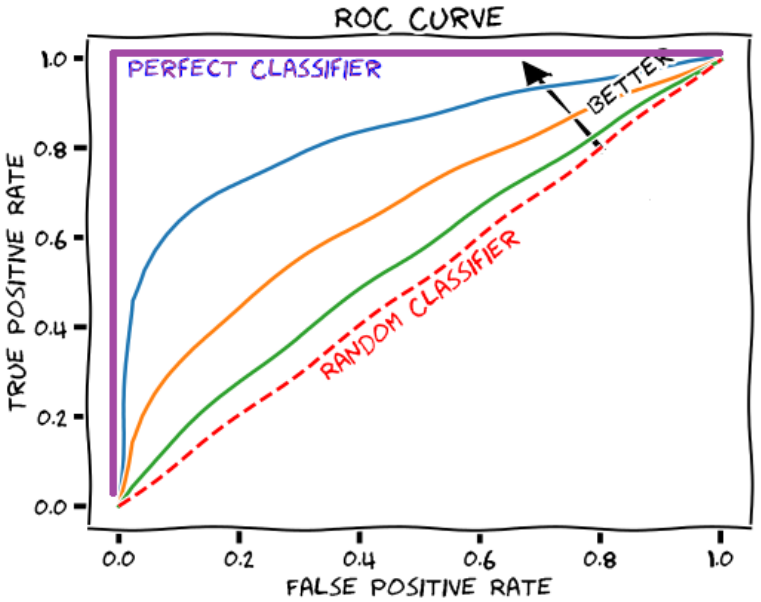
$$Mean\ square\ error : \frac{\sum_{i=1}^{d} (y_i - y_i')^2}{d}$$

$$Relative\ absolute\ error : \frac{\sum_{i=1}^{d} |y_i - y_i'|}{\sum_{i=1}^{d} |y_i - \bar{y}|}$$

$$Relative\ square\ error : \frac{\sum_{i=1}^{d} (y_i - y_i')^2}{\sum_{i=1}^{d} (y_i - \bar{y})^2}$$

Model Selection: ROC Curve

- X: false positive rate
  - f_pos / neg
- Y: true positive rate
  - t_pos / pos
- Area below curve:
  - accuracy, diagonal line: 0.5 accuracy



Model Selection: T-test

- Two models M_1 and M_2, k-fold cross-validation
- $err(M_1)_1, \ldots, err(M_1)_k$ vs. $err(M_2)_1, \ldots, err(M_2)_k$
- Choose model with lower mean error?
- Statistically significant? Or by chance?
- T-test

$$t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{var(M_1 - M_2)/k}}$$

$$var(M_1 - M_2) = \frac{1}{k} \sum_{i=1}^{k} \left[err(M_1)_i - err(M_2)_i - (\overline{err}(M_1) - \overline{err}(M_2)) \right]^2$$

T-test Example

ν	0.90	0.95	0.975	0.99	0.995	0.999
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296

- 10-fold cross-validation
- T-table (e.g., https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm)
- Degree of freedom: v = 10-1 = 9
- Significance level: a = 0.05
- Two-sided test: 1-a/2 = 1-0.05/2 = 0.975
- Check T-table (v=9, 0.975): critical value is 2.262
- Compute t, statistically significant if t > 2.262

Summary: Classification

➢ Supervised learning

- Training set with predefined class labels

➢ Methods

- Decision tree, Bayesian, SVM, neural network, ensemble

➢ Model evaluation, model selection

- Confusion matrix, accuracy, error, ROC curve, T-test