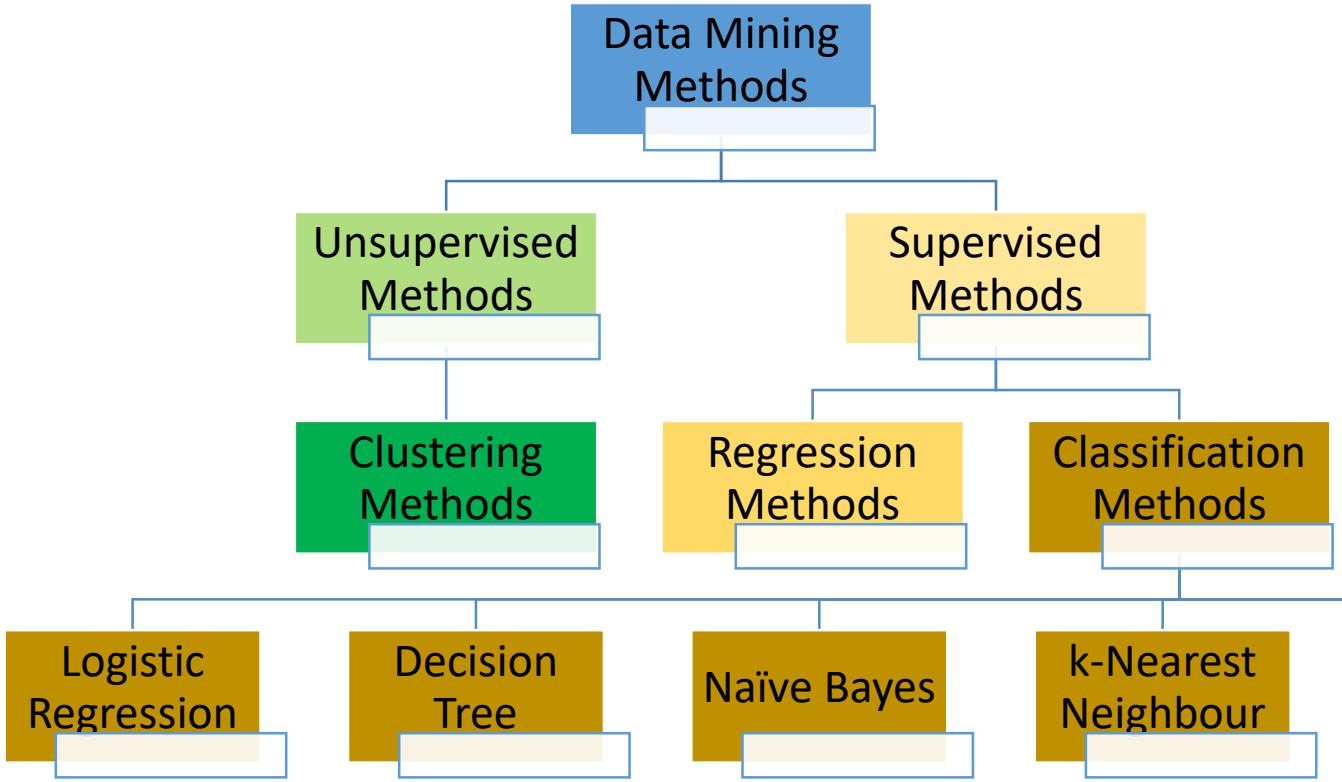
SUPERVISED LEARNING

CLASSIFICATION

Data Mining Tasks in Discovering Knowledge in Data

(Larose, 2005)

Supervised vs Unsupervised Methods



UNSUPERVISED
- no target variable specified
- searches for patterns and structure among all the variable

SUPERVISED
- there is a particular prespecified target variable
- the algorithm may learn from values of the target variable which associated with values of the predictor variables.

(Larose, 2005)

Classification

- In classification task, we **determine the classifier**.
- The classifier **predicts the class of each instance**:
  - if it is correct, that is counted as a ***success***
  - if not, it is an ***error***.

  So, we measure a classifier's performance interms of the error rate.
- What we are interested in is the likely future performance on new data, not the past performance on old data.

(Witten, Frank, & Hall, 2011)

Classification: training data and test data

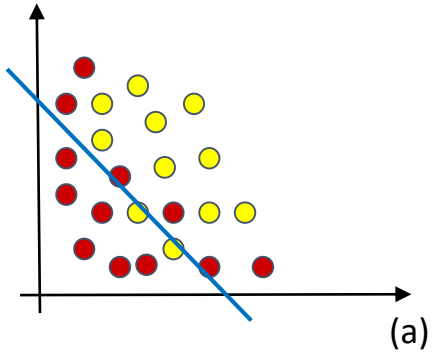
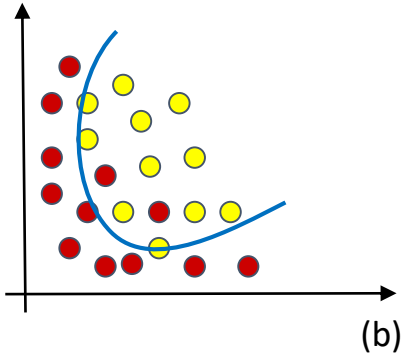
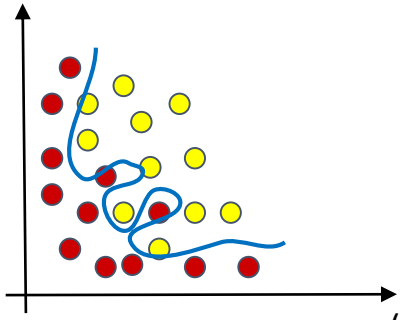
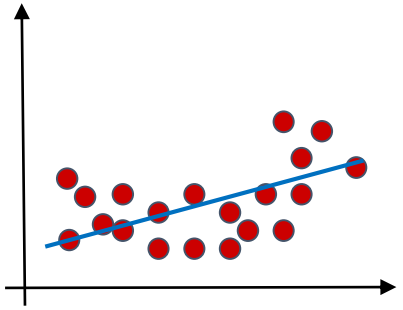
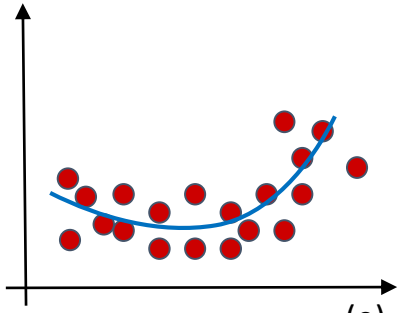
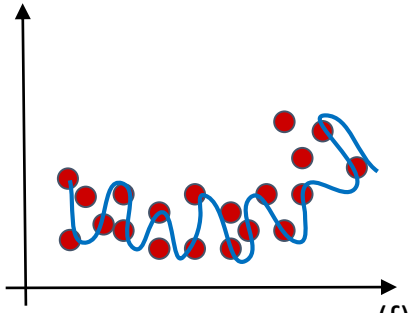
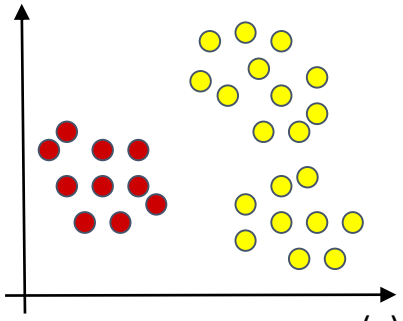
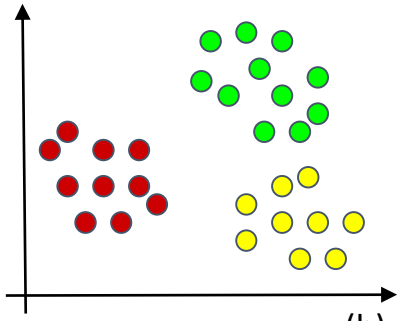
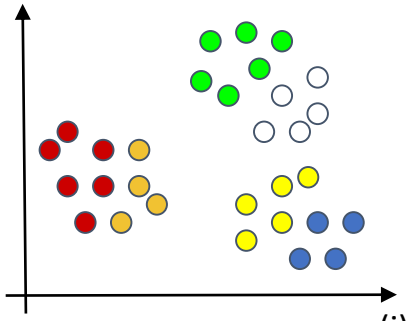
- **Is the error rate on old data likely to be a good indicator of the error rate on new data?** NO—not if the old data was used during the learning process to train the classifier.
- The error rate on the training set is not likely to be a good indicator of future performance. **Why?**
- Because the classifier has been learned from the very same training data, any estimate of performance based on that data will be optimistic, even hopelessly optimistic.
- However, it is often useful to know.
- The independent dataset which is no part in the formation of the classifier is called the **test set**. It is needed to predict the performance of a classifier on new data.

(Witten, Frank, & Hall, 2011)

Classification: training data, validation data, and test data

- In such situations people often talk about three datasets: the training data, the validation data, and the test data.
- The training data is used by one or more learning schemes **to come up** with classifiers.
- The validation data is used **to optimize** parameters of those classifier, or to select a particular one.
- The test data is used **to calculate** the error rate of the final, optimized, method.
- Each of the three sets must be chosen independently.

(Witten, Frank, & Hall, 2011)

	Underfit	Appropriate	Overfit
Classification	 (a)	 (b)	 (c)
Regression	 (d)	 (e)	 (f)
Clustering	 (g)	 (h)	 (i)

Metode Klasifikasi

Naïve Bayes

Metode yang menggunakan Teorema Bayes dengan asumsi independen bersyarat (*conditional independence*).

Decision Tree

Seperti flowchart, terdiri dari node akar yang paling atas, node internal menunjukkan pengujian pada atribut, setiap cabang mewakili hasil pengujian, dan setiap node terminal (leaf node) menunjukkan label kelas.

Support Vector Machine (SVM)

Metode yang memetakan data ke dalam ruang fitur berdimensi tinggi untuk menemukan hyperplane yang optimal

Neural Network (NN)

Metode supervised yang memiliki bentuk seperti jaringan syaraf (neuron), terdiri dari input, hidden, dan output layer. Algoritma dalam NN, misalnya Backpropagation. Multilayer Feed-Forward, dll.

k-Nearest Neighbor (k-NN)

Setiap objek/instance baru dibandingkan dengan yang sudah ada menggunakan metrik jarak, dan *k* instance terdekat digunakan untuk menetapkan kelas ke kelas yang baru.

Regresi Logistik

Metode supervised dengan cara meregresikan respon yang berupa data kategorik dengan prediktor. Respon dapat berupa biner, multinomial, dan ordinal.

dll.

Naïve Bayes

Likelihood

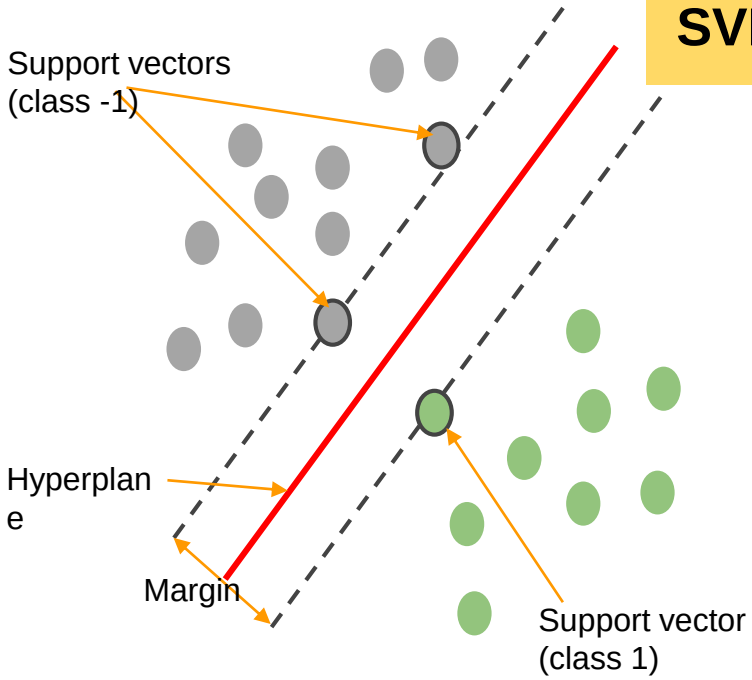
Prior Probability

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$$

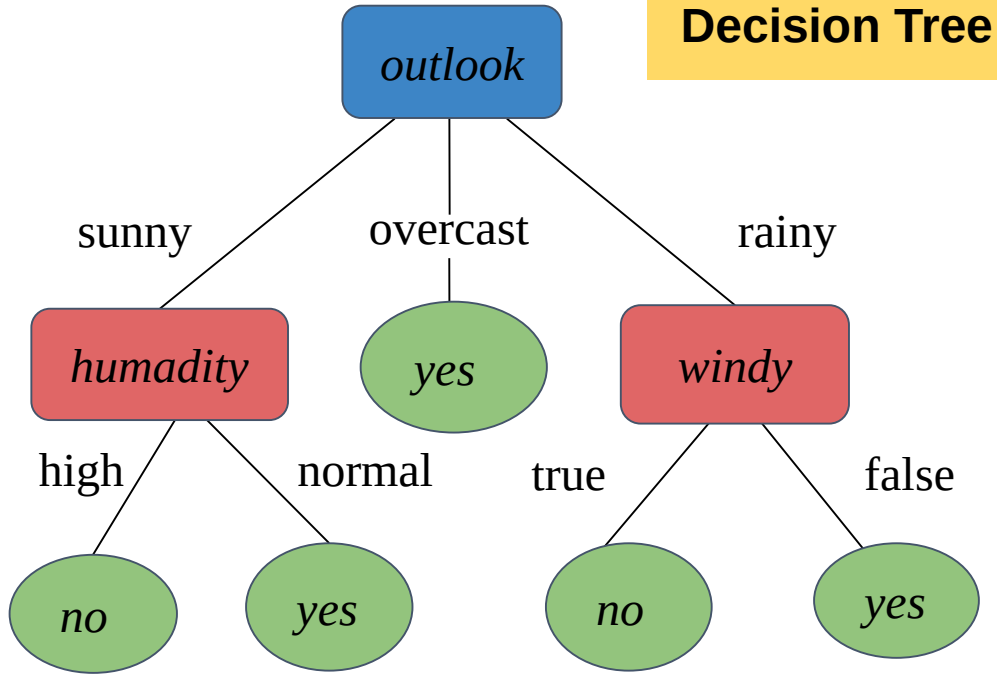
Posterior Probability

Evidence/ feature

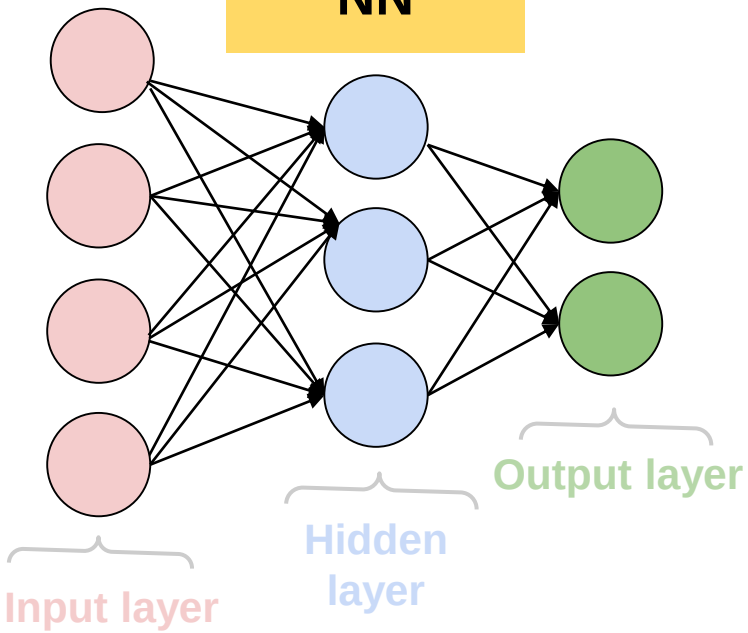
SVM



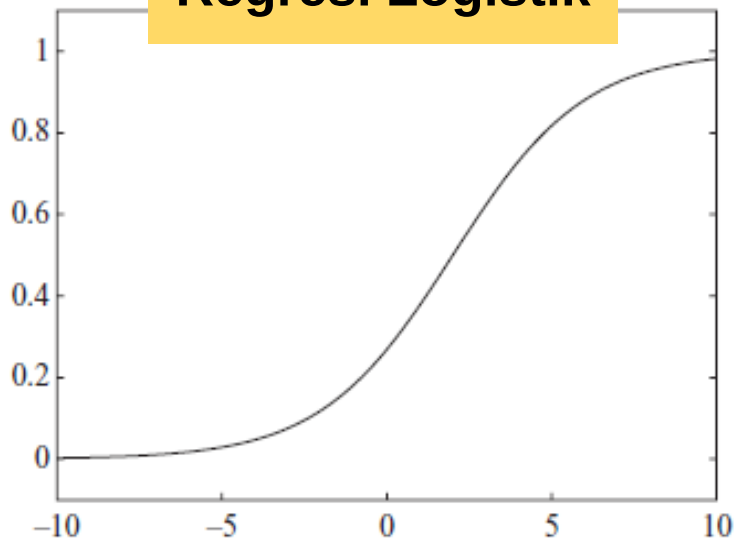
Decision Tree



NN

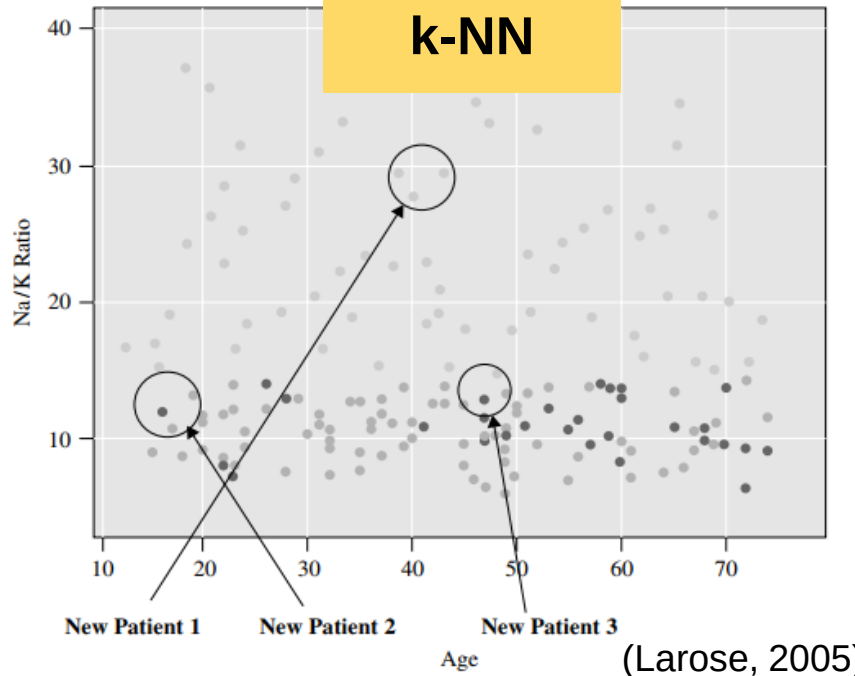


Regresi Logistik



(Witten, Frank, & Hall, 2011)

k-NN



(Larose, 2005)

Model Evaluation: Metrics

- A **confusion matrix** can be used to evaluate a classifier's quality.
- For a two-class problem, it shows the *true positives, true negatives, false positives*, and *false negatives*.
- Measures that assess a classifier's predictive ability include :
  - **accuracy**,
  - **sensitivity** (also known as **recall**),
  - **specificity**,
  - **precision**,
  - *F*, and
  - F_β.
- Reliance on the accuracy measure can be deceiving when the main class of interest is in the minority.

(Han, J., Kamber, M., and Pei, J., 2012)

		Predicted Class	
		yes (pos)	no (neg)
Actual Class	yes (pos)	TP	FN
	no (neg)	FP	TN

True positives: the positive tuples that were correctly labeled by the classifier.
True negatives : These are the negative tuples that were correctly labeled by the classifier.
False positives : These are the negative tuples that were incorrectly labeled as positive.
False negatives : These are the positive tuples that were mislabeled as negative

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Model Evaluation: Partition

- Construction and evaluation of a classifier require partitioning labeled data into a training set and a test set.
- Holdout, random sampling, cross-validation, and bootstrapping are typical methods used for such partitioning.

(Han, J., Kamber, M., and Pei, J., 2012).

Referensi

- Han, J., Kamber, M., and Pei, J. (2012). *Data Mining Concepts and Techniques 3rd Edition.* USA: Morgan Kaufmann.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques 3rd Edition.* USA: Morgan Kaufmann.
- Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining,* New Jersey:John Wiley & Sons, Inc.

Terima Kasih