Intro to Classification and Regression

Classification and Regression

- Classification:
  - predicts categorical class labels
  - Constructs classification models based on training data and uses the models in classifying new data
- Regression:
  - models continuous-valued functions, i.e., predicts unknown or missing numeric values
- Example Applications
  - credit approval- classify loan application by their likelihood of defaulting on payments
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

Classification Applications

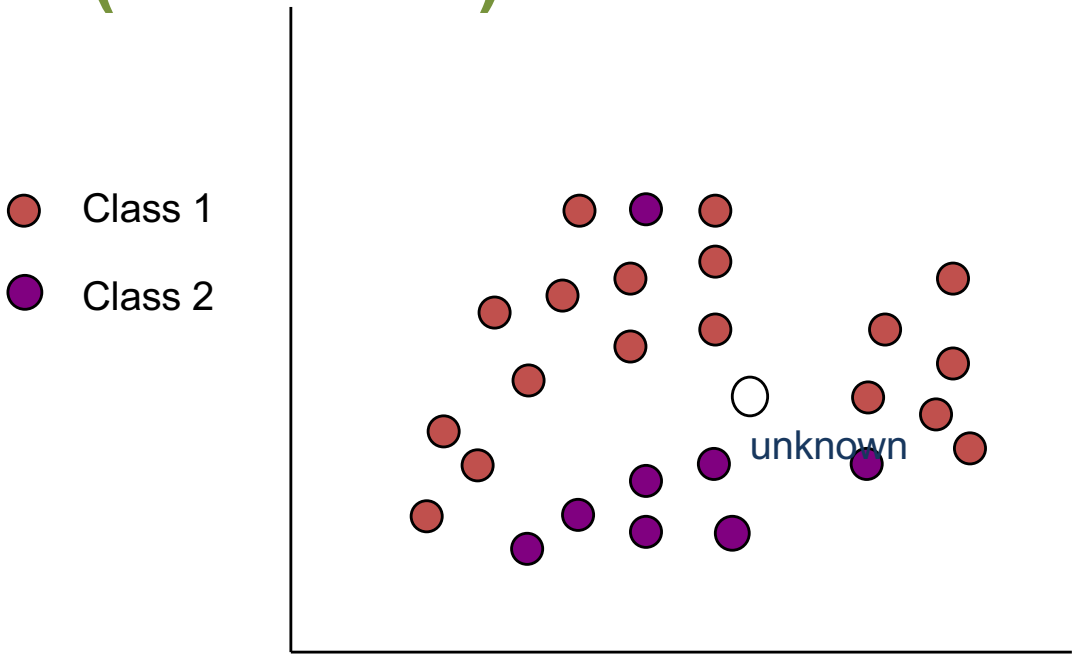
- Example Applications (continued)
  - Image processing : interpretation of digital images in radiology, recognizing 3-D objects, outdoor image segmentation
  - Language processing : text classification
  - Software development : estimate the development effort of a given software module
  - Pharmacology: drug analysis
  - Molecular biology : analyzing amino acid sequences
  - Medicine : cardiology, analyzing sudden infant death syndrome, diagnosing thyroid disorder
  - Manufacturing : classify equipment malfunctions by their cause

Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction: training set
  - The model is represented as classification rules, decision trees, mathematical formulae, neural networks, or an ensemble of these
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The set of tuples used for testing the performance of the model: test data
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur

Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances



Many approaches: Regression, Decision Trees, Nearest Neighbor, Support Vector Machines, Neural Networks,

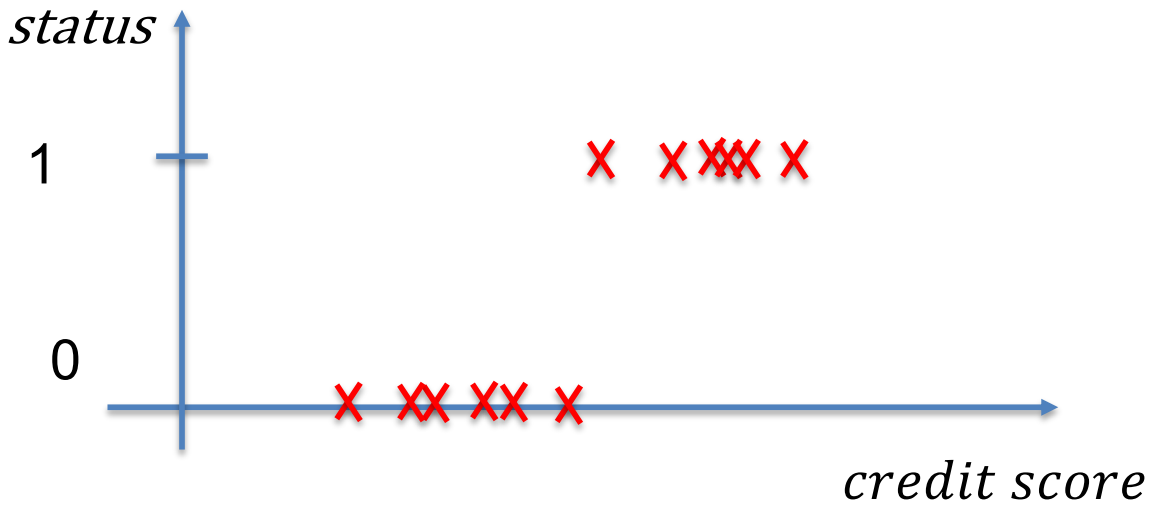
...

Classification Problem

Loan approval problem with a single variable

x_1: credit score (FICO score)
y: 1-approve, 0-deny

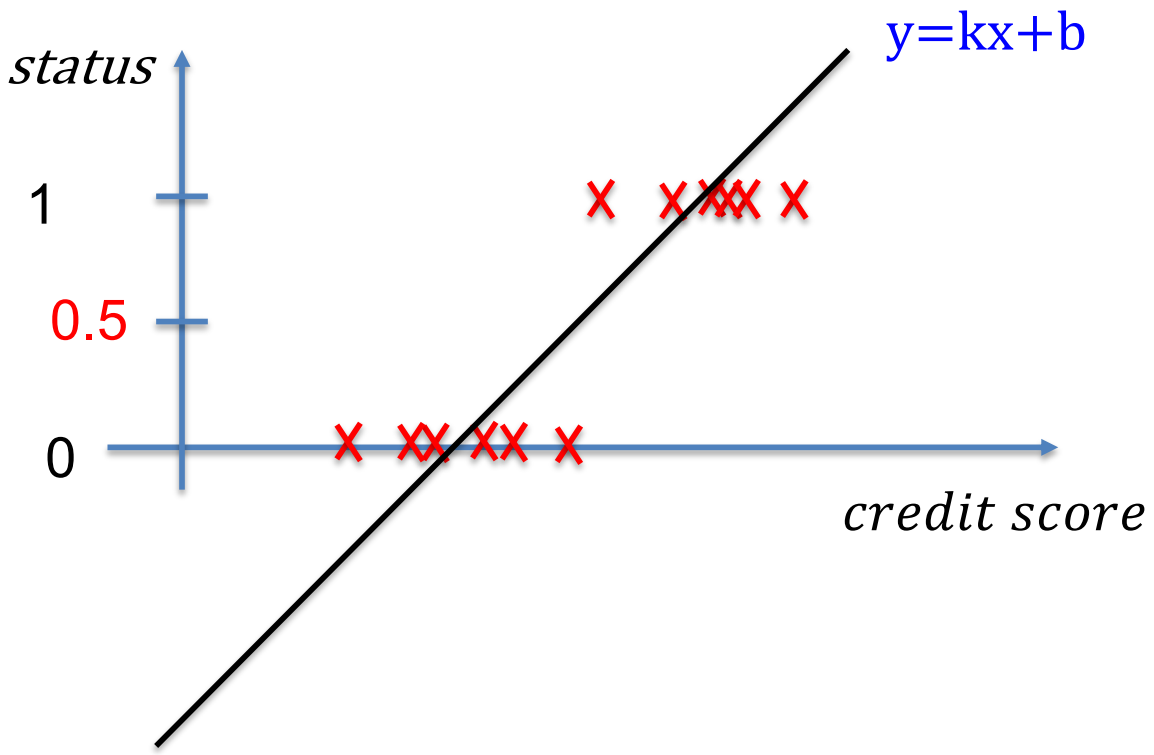
Credit Score	Loan Status
750	1
725	0
700	0
650	0
726	1
645	0
800	1
…	…



Classification Problem

Loan approval problem with a single variable

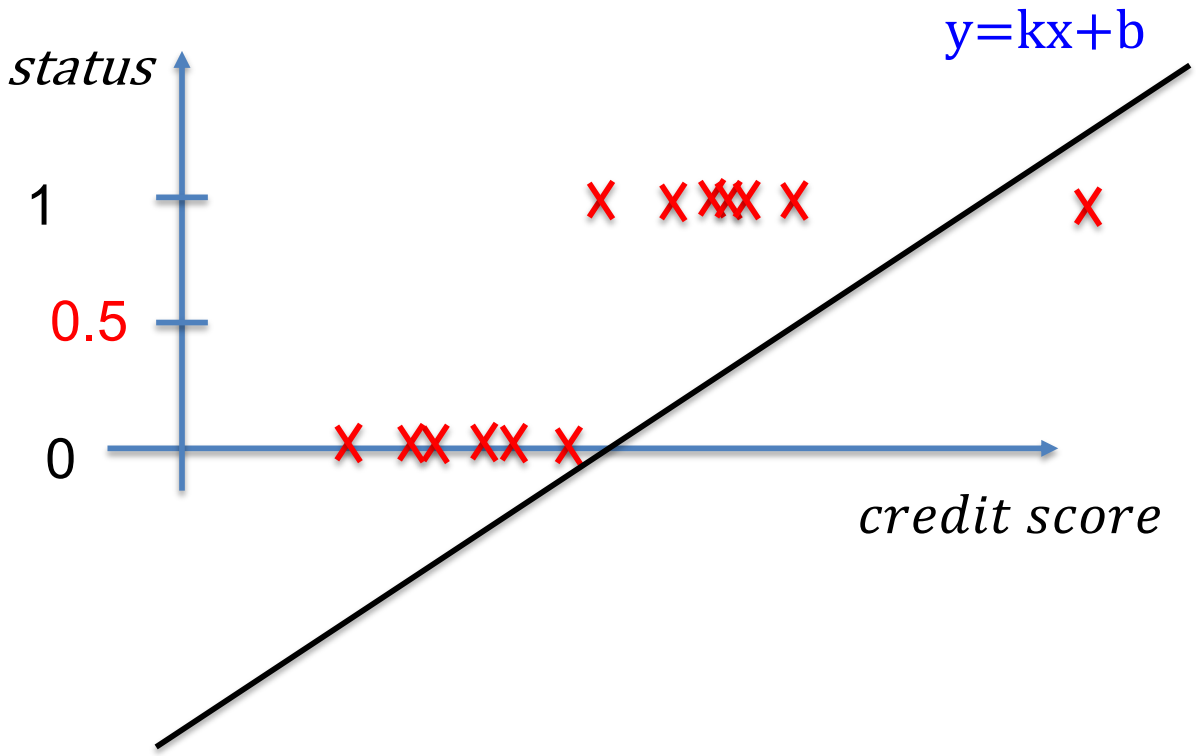
Credit Score	Loan Status
750	1
725	0
700	0
650	0
726	1
645	0
800	1
…	…



Classification Problem

Loan approval problem with a single variable

Credit Score	Loan Status
750	1
725	0
700	0
650	0
726	1
645	0
800	1
…	…



Classification Problem

Loan approval problem

x_1: credit score (FICO score)
x_2: income
(may include other features)

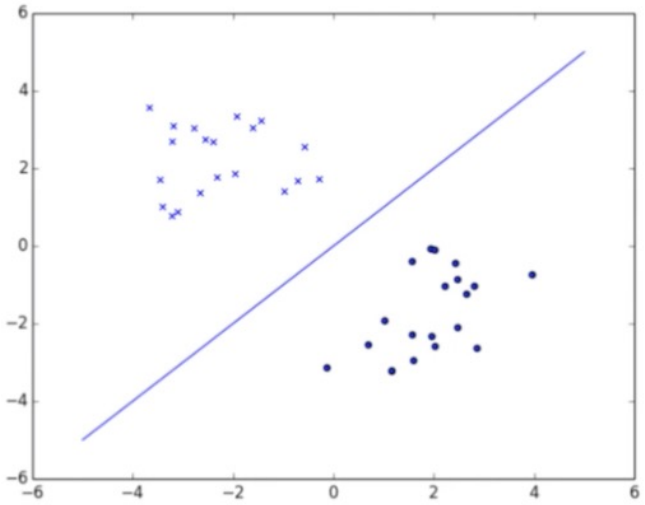
y: 1-approve, 0-deny

Training Data

Credit Score	Income	Loan Status
750	113000	1
725	26000	0
700	54000	0
650	45000	0
726	89500	1
645	78500	0
800	87050	1
...	...	...

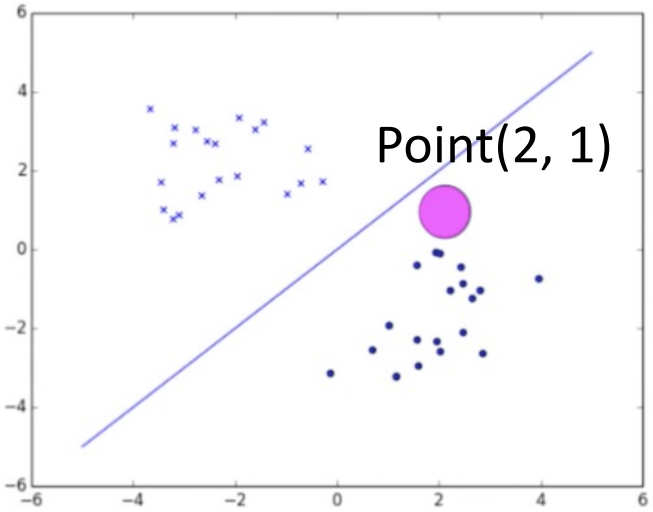
Test data:
for a new applicant with credit score 715 and
income 68500, will the loan application be approved?

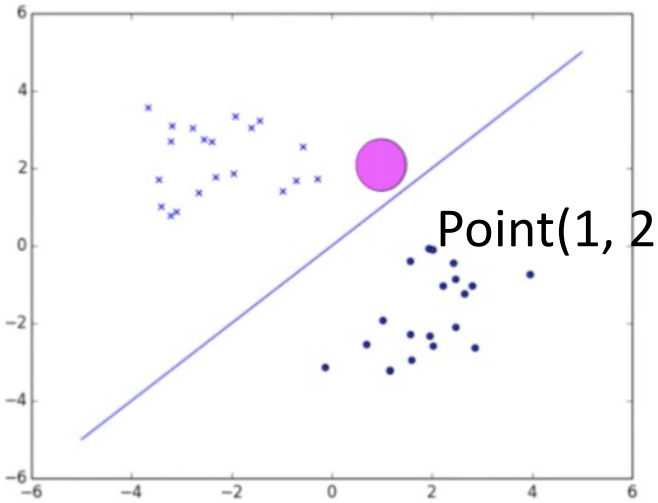
Linear Regression



$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$





Regression

- $h_\theta(\mathrm{x}) = \theta_0 + \theta_1 \mathrm{x}_1 + \theta_2 \mathrm{x}_2$, h() is a linear combination of the components of x
  - In vector form : $h_\theta(\mathrm{x}) = \theta^{\mathsf{T}}\mathrm{x}$
- The class separating function:
  - In 2-dimensions: a line
  - In 3-dimensions: a plane
  - In >3 dimension: hyperplane