Contents

1 Module 02: Classification

- Definition: putting things into groups

1.1 Types of classification models

1. Number of groups
2. Number of dimensions
   - Can 1 dimension be sufficient to classify?
3. Soft vs hard classifiers (is it 100% error free?)

1.2 Definition of bad classification

- Cost: is one type of mistake worse than the other?

1.3 Examples

1.3.1 Loan payment (Income vs credit score)

- Plot lines and find one that can separate default vs non-default.
- How do we know the right lines are drawn?
- We want to be as conservative as possible (less error prone)

1.4 Data terminology

1. Row = data point
2. Column = dimension, attribute, feature, predictor, covariate
   (a) Special column = response, outcome

1.5 Data types

1. Structured data
   (a) Quantitative
      - Numbers with meaning

   (b) Categorical
      - Numbers without meaning

   (c) Binary data (subset of categorical)
   (d) Unrelated data
   (e) Time series data
2. Unstructured
   (a) Text data

1.6 Support vector machines

- **Supervised** method (algorithm uses known results when training)
- Terminology
  - m = number of data points
  - n = number of attributes
  - x_{ij} = j attribute of i data point
    * e.g. x_{51} = credit score of person 5; x_{52} = income of person 5
  - y_i = response of data point i
    * e.g. 1 if data point is group 1
    * -1 if data point is group 2
  - Line: $a_1x_1 + a_2x_2 + \ldots + a_nx_n + a_0 = 0$
  - Note the intercept a_0

- In general: $\sum_{j=1}^{n} a_j x_j + a_0 = 0$
- Separation problem: get max distance between lines
- $\frac{2}{\sqrt(\sum_j (a_j)^2)}$
- i.e. $\text{Min}_{\text{a}_0 \ \ldots \ \text{a}_\text{n}}$: $\sum_{j=1}^{n} (a_j)^2$
- Subject to constraints

1.6.1 When not possible to get full separation

- Then we minimize error
- There's a trade-off between margin and error
- Error for data point is:

$$\max\{0, 1 - (\sum_{j=1}^{n} a_j x_{ij} + a_0) y_i\}$$

- We multiply margin by λ and assign its importance of **margin** vs error that way.
- Full equation is: #TODO