Linear Models for Classification: Introduction

Sargur N. Srihari

University at Buffalo, State University of New York

USA

Topics

- Regression vs Classification
- Linear Classification Models
- Converting probabilistic regression output to classification output
- Three classes of classification models

Regression vs Classification

- In *Regression* we assign input vector x to one or more continuous target variables t
  - Linear regression has simple analytical and computational properties

- In *Classification* we assign input vector x to one of K discrete classes $C_k,\ k = 1, \ldots ,K.$
  - We discuss here linear models for Classification
  - Ordinal Regression is a form of classification where discrete classes have an ordering
    - E.g., relevance score regression

Linear Classification Models

- Common classification scenario: classes considered disjoint

  – Each input assigned to only one class

- Input space divided into decision regions

- Decision surfaces are linear functions of input x

  – Defined by $(D-1)$ dimensional hyperplanes within D dim. input space
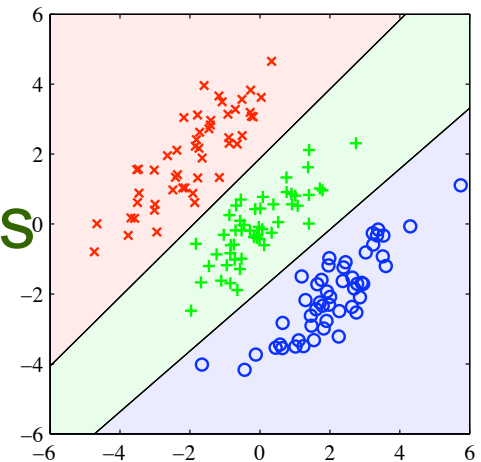
Straight line is 1-D in 2-D
A plane is 2-D in 3-D

Data sets whose classes can be separated exactly by linear decision surfaces are said to be Linearly separable

Representing the target in Classification

- In *regression* target variable t is a real number (or vector of real numbers **t**) which we wish to predict

- In *classification* there are various ways of using target values to represent class labels, depending on whether
  - Model is probabilistic
  - Model is non-probabilistic

Representing Class in Probabilistic Model

- ## Two class: Binary representation is convenient
  - Discrete $t \in \{0, 1\}$, $t = 1$ represents C_1,
    $t = 0$ means class C_2
    - Can interpret value of t as probability that class is C_1
    - Probabilities taking only extreme values of 0 and 1

- ## For $K > 2$ - Use a 1-of-K coding scheme.
  - $\mathbf{t}$ is a vector of length K
    - Eg. if $K = 5$, a pattern of class 2 has $\mathbf{t} = (0, 1, 0, 0, 0)^T$
    - Value of t_k interpreted as probability of class C_k
      - If t_k assume real values then we allow different class probabilities

Representing Class: Nonprobabilistic

- For non-probabilistic models, e.g,
nearest neighbor
  - other choices of target variable
representation used

Two Approaches to Classification

1. ## Discriminant function

   – Directly assign $\mathbf{x}$ to a specific class

      • E.g., Fisher's Linear Disc, Perceptron

2. ## Probabilistic Models

   1. Model $p(C_k/\mathbf{x})$ in *inference* stage (direct or $p(\mathbf{x}|C_k)$)

   2. Use it to make *optimal* decisions

Separating Inference from Decision is better:

      • Minimize risk (loss function can change in financial app)

      • Reject option (minimize expected loss)

      • Compensate for unbalanced data

         –   use modified balanced data & scale by class fractions

      • Combine models

Two Prob. Models for Classification

- Model $p(C_k/\mathbf{x})$ in an *inference* stage and use it to make optimal decisions

- Two approaches to computing the $p(C_k/\mathbf{x})$

  - Generative

    - Model class conditional densities by $p(\mathbf{x}|C_k)$ together with prior probabilities $p(C_k)$
    - Then use Bayes rule to compute posterior

      $$p(C_k|\mathbf{x}) = p(\mathbf{x}|C_k)p(C_k)/p(\mathbf{x})$$

  - Discriminative

    - Directly model conditional probabilities $p(C_k|\mathbf{x})$

Converting Linear Regression Model to Linear Classification

- Linear Regression model $y(\mathrm{x},\mathrm{w})$ is a linear function of parameters w

- In simplest case also a linear function of variables x

  – Thus has the form $y(\mathrm{x})=\mathrm{w}^{\mathrm{T}}\mathrm{x}+w_0$

- For classification we wish to obtain a discrete output or posterior probabilities in range $(0,1)$

- Use *generalized linear model*

$$y(\mathrm{x})=f\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}+w_0\right)$$

Generalized Linear Model

$$y(\mathrm{x})=f\left(\mathrm{w^Tx}+w_0\right)$$

$f\left(.\right)$ is known as the activation function

- The decision surfaces correspond to
$y(\mathrm{x})=constant$ or $\mathrm{w}^T\mathrm{x}+w_0=constant$

  – So decision boundaries are linear in feature space x even if $f\left(.\right)$ is nonlinear, e.g., a step function

  – Hence called generalized linear model

- However no longer linear in parameter space w due to presence of $f\left(.\right)$

  – Leads to more complex models for classification than regression

11
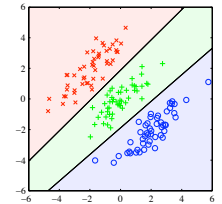
Overview of Linear Classifiers

1. ## Discriminant Functions
   - Two class and Multi class
   - Least squares for classification
   - Fisher's linear discriminant
   - Perceptron algorithm

2. ## Probabilistic Generative Models
   - Continuous inputs and max likelihood
   - Discrete inputs, Exponential Family

3. ## Probabilistic Discriminative Models
   - Logistic regression for single and multi class
   - Laplace approximation
   - Bayesian logistic regression

12