Linear Models for Classification: Introduction

Sargur N. Srihari
University at Buffalo, State University of New York
USA

Topics

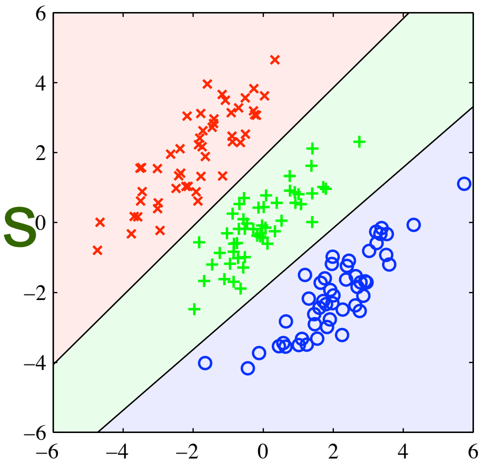
- Regression vs Classification
- Linear Classification Models
- Converting probabilistic regression output to classification output
- Three classes of classification models

Regression vs Classification

- In *Regression* we assign input vector $\boldsymbol{x}$ to one or more continuous target variables t
  - Linear regression has simple analytical and computational properties
- In *Classification* we assign input vector $\boldsymbol{x}$ 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 $\boldsymbol{x}$
  - Defined by $(D-1)$ dimensional hyperplanes within D dim. input space



Straight line is1-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 $\mathbf{t}$) 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 $1\text{-of-}K$ coding scheme.
  - $\mathbf{t}$ is a vector of length K
    - Eg. if $K = 5$, a pattern of class 2 has $\mathbf{t} = (0, 1, 0, 0, 0)^{\mathrm{T}}$
    - 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 $\boldsymbol{x}$ to a specific class
     - E.g., Fisher's Linear Disc, Perceptron
2. Probabilistic Models
   1. Model $p(C_k|\boldsymbol{x})$ in *inference* stage (direct or $p(\boldsymbol{x}|C_k)$)
   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

Probabilistic Models: Generative/Discriminative

- Model $p(C_k|\boldsymbol{x})$ in an *inference* stage and use it to make optimal decisions
- Two approaches to computing the $p(C_k|\boldsymbol{x})$
  - Generative
    - Model class conditional densities by $p(\boldsymbol{x}|C_k)$ together with prior probabilities $p(C_k)$
    - Then use Bayes rule to compute posterior

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

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

From Regression to Classification

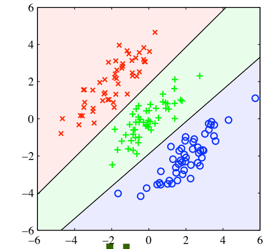
- Linear Regression model $y(\boldsymbol{x},\boldsymbol{w})$ is a linear function of parameters $\boldsymbol{w}$
  - In simple case model is also a linear function of $\boldsymbol{x}$
    - Thus has the form $y(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+w_0$ where y is a real no.
- For classification we need need to predict class labels or posterior probabilities in range $(0,1)$
  - For this, we use a generalization where we transform the linear function of w using a nonlinear function $f(.)$, so that

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

- $f(.)$ is known as an *activation function*
- Whereas its inverse is called a *link function* in statistics
  - link function provides relationship between the linear predictor and the mean of the distribution function

Decision surface of linear classifier

- Decision surfaces of $y(\boldsymbol{x})=f\,(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+w_0)$ correspond to $y(\boldsymbol{x})=\text{constant}$ or $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+w_0=\text{constant}$



  – Surfaces are linear in $\boldsymbol{x}$ even if $f\,(.)$ is nonlinear
    - For this reason they are called *generalized linear models*

  – However no longer linear in parameters $\boldsymbol{w}$ due to presence of $f\,(.)$, therefore:
    – More complex models for classification than regression

- Linear classification algorithms we discuss are applicable even if we transform $\boldsymbol{x}$ using a vector of basis functions $\boldsymbol{\phi}(\boldsymbol{x})$

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