Linear Models for Classification: Overview

Sargur N. Srihari

University at Buffalo, State University of New York

USA

Topics in Linear Models for Classification

- Overview

1. Discriminant Functions
2. Probabilistic Generative Models
3. Probabilistic Discriminative Models
4. The Laplace Approximation

Topics in Overview

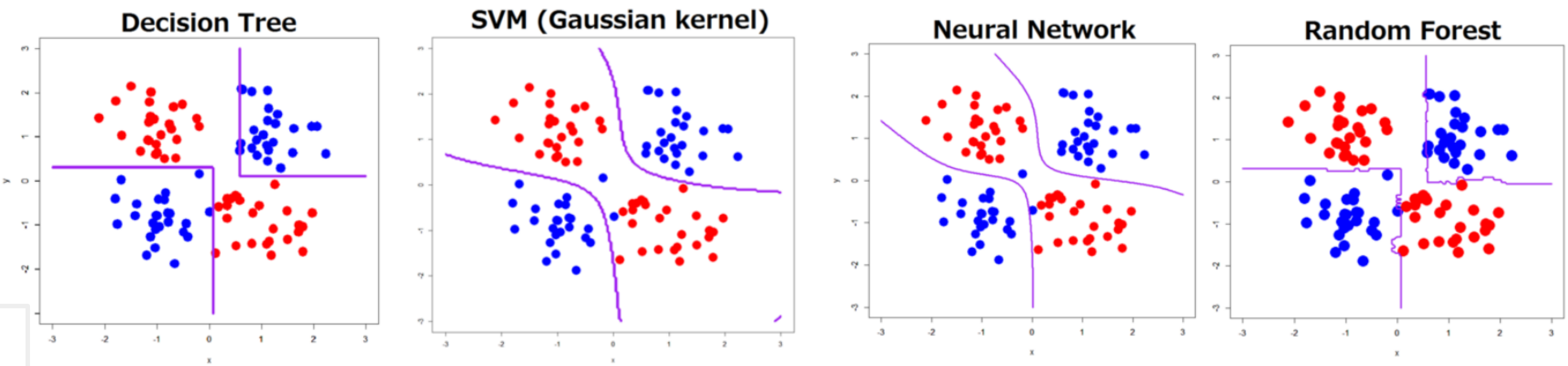
1. Regression vs Classification
2. Linear Classification Models
3. Converting probabilistic regression output to classification output
4. Three classes of classification models

Regression vs Classification

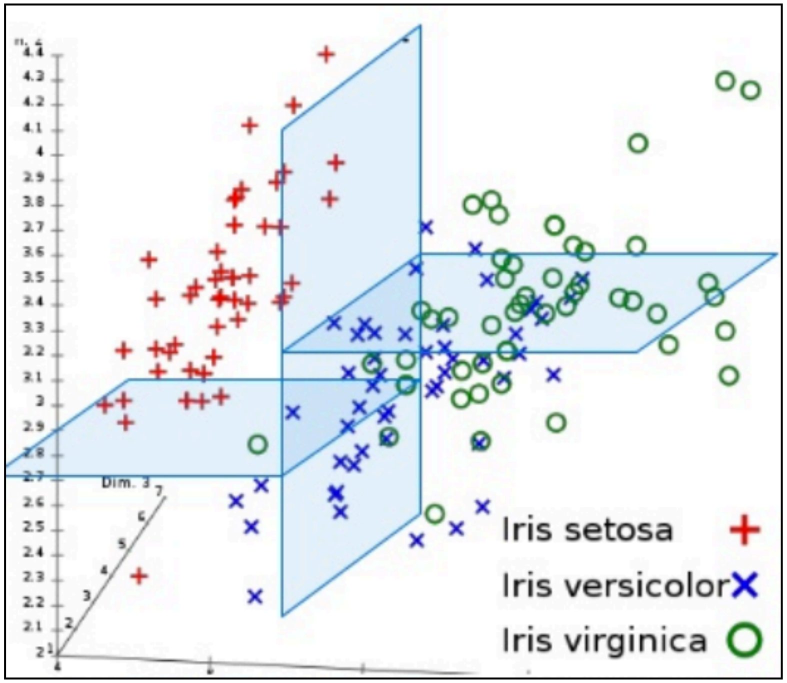
- In *Regression* we assign input vector $\boldsymbol{x}$ to one or more continuous target variables t
  - Linear regression has simple analytical and computational properties
- In *Classification* we assign input vector $\boldsymbol{x}$ to one of K discrete classes C_k, $k = 1, \ldots, K$
  - Common classification scenario: classes considered disjoint
    - Each input assigned to only one class
  - Input space is thereby divided into decision regions

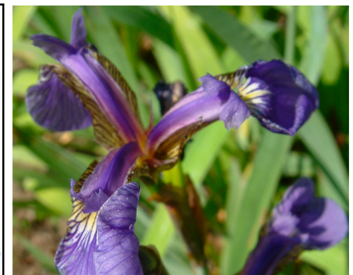
Boundaries of decision regions

Boundaries are called *decision boundaries* or *decision surfaces*



Decision tree with linear boundaries



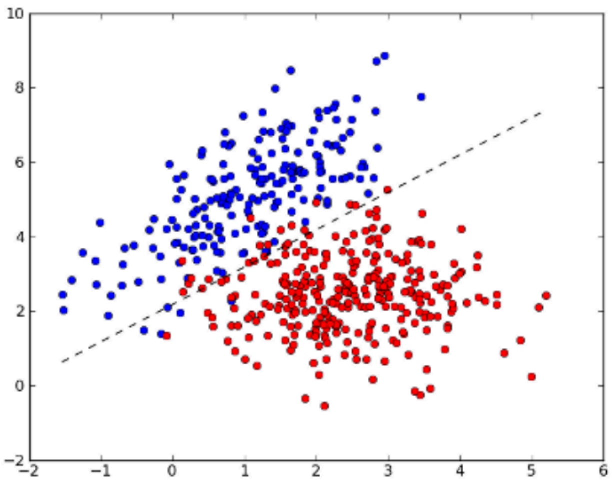


Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.3	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
7.0	3.2	4.7	1.4	*I. versicolor*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.5	2.8	4.6	1.5	*I. versicolor*
5.7	2.8	4.5	1.3	*I. versicolor*
6.3	3.3	4.7	1.6	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
5.8	2.7	5.1	1.9	*I. virginica*
7.1	3.0	5.9	2.1	*I. virginica*
6.3	2.9	5.6	1.8	*I. virginica*
6.5	3.0	5.8	2.2	*I. virginica*
7.6	3.0	6.6	2.1	*I. virginica*
4.9	2.5	4.5	1.7	*I. virginica*
7.3	2.9	6.3	1.8	*I. virginica*

Linear Classification Models

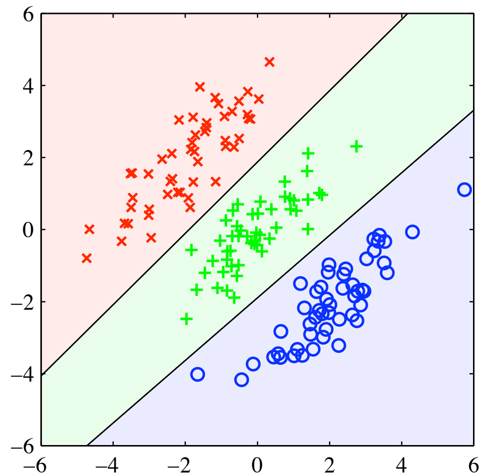
- Decision surfaces are linear functions of input $\boldsymbol{x}$
  - Defined by $(D-1)$ dimensional hyperplanes within D dimensional input space

$K=2$



Not linearly separable

$K=3$



Linearly separable

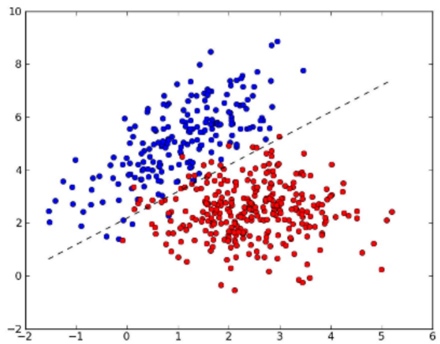
Straight line is 1-D decision boundary in 2-D space

A plane is 2-D surface in 3-D space

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-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

Three Approaches to Classification

1. Discriminant function
   - Directly assign $\boldsymbol{x}$ to a specific class
     - E.g., Linear discriminant, Fisher Linear Disc, Perceptron
2. Probabilistic Models (2)
   1. Discriminative approach
      - Model $p(C_k|\boldsymbol{x})$ in *inference* stage (direct)
      - Use it to make *optimal* decisions
      - E.g., Logistic Regression
   2. Generative approach
      - Model class-conditional density $p(\boldsymbol{x}|C_k)$
      - Together with $p(C_k)$ use Bayes rule to compute posterior
      - E.g., Naiive Bayes classifier

Separating Inference from Decision

- 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
- Approaches to computing the $p(C_k|\boldsymbol{x})$
  1. 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

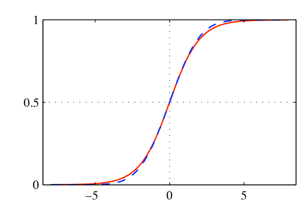
$$p(C_k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_k)p(C_k)}{p(\boldsymbol{x})}$$

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

From Regression to Classification

- Linear Regression: model prediction $y(\boldsymbol{x},\boldsymbol{w})$ 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.
- 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 $\boldsymbol{w}$ using a nonlinear function $f(.)$, so that

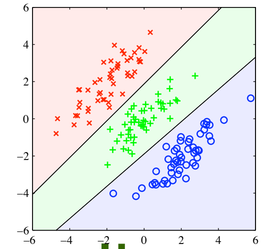
$$y(\boldsymbol{x})=f\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+w_0\right)$$



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

Well-known Probabilistic Models