Probabilistic Discriminative Models: Fixed Basis Functions

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

Topics in Linear Classification using Probabilistic Discriminative Models

- Distinction between Generative vs Discriminative
- 1. Fixed basis functions
- 2. Logistic Regression (two-class)
- 3. Iterative Reweighted Least Squares (IRLS)
- 4. Multiclass Logistic Regression
- 5. Probit Regression
- 6. Canonical Link Functions

- Two-class classification: Posterior of class C_1 : $p(C_1|x)$ can be written as a logistic sigmoid operating on linear function of x, i.e., $\sigma(w^Tx + w_0)$, for wide choice of forms for $p(x|C_k)$
 - For Gaussians with same covariance matrix $\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 \mu_2)$ $\mathbf{w}_0 = -\frac{1}{2}\mu_1^T \mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2^T \mathbf{\Sigma}^{-1}\mu_2 + \ln \frac{p(C_1)}{p(C_2)}$
 - For Discrete binary features x_i
 - which are linear functions of features

$$\begin{split} a_k(\boldsymbol{x}) &= \ln(p(\boldsymbol{x} \mid \boldsymbol{C}_k) p(\boldsymbol{C}_k)) \\ &= \sum_{i=1}^{D} \left\{ x_i \ln \boldsymbol{\mu}_{ki} + (1-x_i) \ln(1-\boldsymbol{\mu}_{ki} \right\} + \ln p(\boldsymbol{C}_k) \end{split}$$

- Multiclass case:
 - Posterior probability of class C_k i.e.,

$$\begin{aligned} p(C_k \mid \boldsymbol{x}) &= \frac{p(\boldsymbol{x} \mid C_k)p(C_k)}{\sum_j p(\boldsymbol{x} \mid C_j)p(C_j)} & \text{Normalized} \\ &= \frac{\exp(a_k)}{\sum_j \exp(a_j)} & \text{Or Softmax} \end{aligned}$$

- $p(C_k|x)$ is given by a softmax transformation of a linear function of x
- MLE used to get parameters of $p(x|C_k)$ and $p(C_k)$
 - This indirect approach to find parameters of a generalized linear model is called generative
 - Since we can use such a model to generate synthetic data from marginal p(x)

Probabilistic Discriminative Models

An alternative approach

- Use the functional form of the generalized linear model explicitly
- Determine its parameters using maximum likelihood
- There is an efficient algorithm to find such solutions known as Iterative reweighted least squares (IRLS)

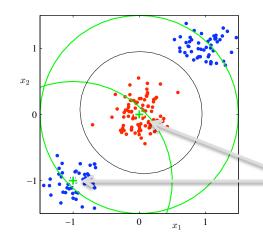
Advantages

- Fewer adaptive parameters
- Improved predictive performance when $p(x|C_k)$ assumptions are poor approximations

Fixed Basis Functions

Although we use linear classification models Linear-separability in feature space does not imply linear-separability in input space

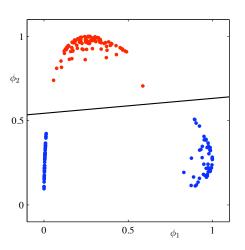
Original Input Space (x_1,x_2)



not linearly separable

Nonlinear transformation of inputs using vector of basis functions $\varphi(x)$

Basis functions are Gaussian with centers Shown by crosses and Green contours Feature Space (ϕ_1, ϕ_2)



linearly separable

Basis functions with increased dimensionality is often used

Limitation of Fixed Basis Functions

- Nonlinear transformations cannot remove overlap between classes
 - They can even increase the overlap!
 - Still fixed nonlinear basis functions play an important role
- One solution: basis functions that adapt to data
 - SVMs use basis functions centered on the data points and select a fixed subset of them