

Linear Classification: Probabilistic Generative Models

Sargur N. Srihari

University at Buffalo, State University of New York
USA

Linear Classification using Probabilistic Generative Models

- Topics
 1. Overview (Generative vs Discriminative)
 2. Bayes Classifier
 - using Logistic Sigmoid and Softmax
 3. Continuous inputs
 - Gaussian Distributed Class-conditionals
 - Parameter Estimation
 4. Discrete Features
 5. Exponential Family

Overview of Methods for Classification

1. Generative Models (Two-step)

1. Infer class-conditional densities $p(\mathbf{x} | C_k)$ and priors $p(C_k)$
2. Use Bayes theorem to determine posterior probabilities

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

2. Discriminative Models (One-step)

- Directly infer posterior probabilities $p(C_k | \mathbf{x})$
- Decision Theory
 - In both cases use decision theory to assign each new \mathbf{x} to a class

Generative Model

- Model class conditionals $p(\mathbf{x} | C_k)$, priors $p(C_k)$
- Compute posteriors $p(C_k | \mathbf{x})$ from Bayes theorem
- Two class Case
- Posterior for class C_1 is

$$p(C_1 | \mathbf{x}) = \frac{p(\mathbf{x} | C_1)p(C_1)}{p(\mathbf{x} | C_1)p(C_1) + p(\mathbf{x} | C_2)p(C_2)}$$

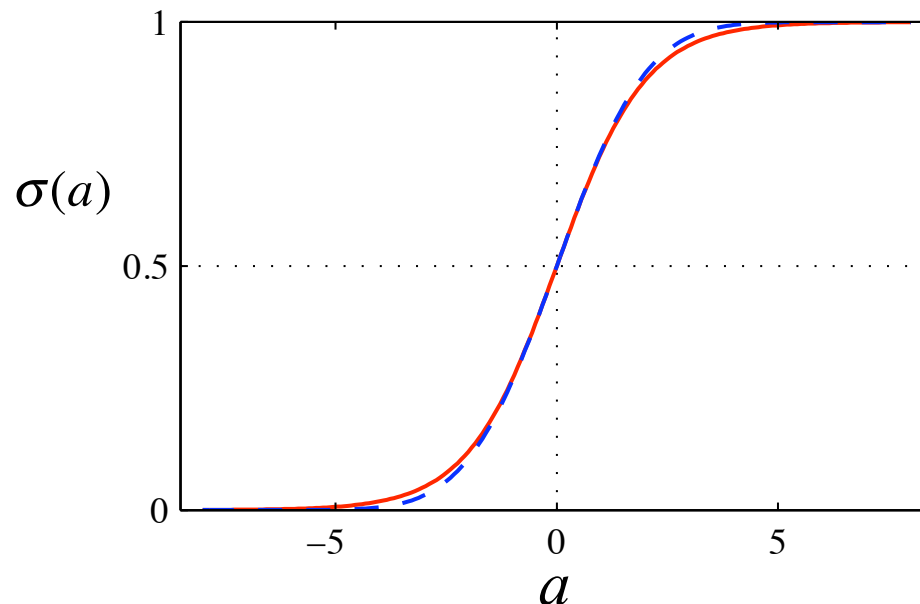
$$= \frac{1}{1 + \exp(-a)} = \sigma(a) \quad \text{where} \quad a = \ln \frac{p(\mathbf{x} | C_1)p(C_1)}{p(\mathbf{x} | C_2)p(C_2)}$$

Since

$$p(\mathbf{x}) = \sum_i p(\mathbf{x}, C_i) = \sum_i p(\mathbf{x} | C_i)p(C_i)$$

LLR with
Bayes odds

Logistic Sigmoid Function



$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

Property : $\sigma(-a) = 1 - \sigma(a)$

Inverse : $a = \ln\left(\frac{\sigma}{1 - \sigma}\right)$

If $\sigma(a) = P(C_1 | \mathbf{x})$ then

Inverse represents

$$\ln[p(C_1 | \mathbf{x}) / p(C_2 | \mathbf{x})]$$

Log ratio of
probabilities
called logit or log
odds

Sigmoid: “S”-shaped or squashing function

maps real $a \in (-\infty, +\infty)$ to finite $(0,1)$
interval

Note: Dotted line is scaled probit function
cdf of a zero-mean unit variance Gaussian

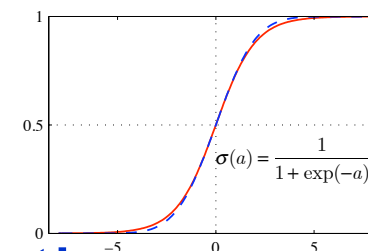
Generalizations and Special Cases

- More than 2 classes
- Gaussian Distribution of x
- Discrete Features
- Exponential Family

Softmax: Generalization of logistic sigmoid

- For $K=2$ we used logistic sigmoid

- $p(C_1|\mathbf{x}) = \sigma(a)$ where $a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$ Log ratio of probabilities



- For $K > 2$, we can use its generalization

$$p(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k)p(C_k)}{\sum_j p(\mathbf{x} | C_j)p(C_j)}$$

$$= \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

If $K=2$ this reduces to a sigmoid

$$\begin{aligned} p(C_1|\mathbf{x}) &= \exp(a_1) / [\exp(a_1) + \exp(a_2)] \\ &= 1 / [1 + \exp(a_2 - a_1)] \\ &= 1 / [1 + \exp(\ln p(\mathbf{x}|C_2)p(C_2) - \ln p(\mathbf{x}|C_1)p(C_1))] \\ &= 1 / [1 + p(\mathbf{x}|C_2)p(C_2) / p(\mathbf{x}|C_1)p(C_1)] \\ &= 1 / [1 + \exp(-a)] \text{ where } a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} \end{aligned}$$

- Quantities a_k are defined by $a_k = \ln p(\mathbf{x}|C_k)p(C_k)$
- Known as the *soft-max* function
 - Since it is a smoothed max function
 - If $a_k \gg a_j$ for all $j \neq k$ then $p(C_k|\mathbf{x}) = 1$ and 0 for rest
 - A general technique for finding max of several a_k

Specific forms of class-conditionals

- We will next see that linear classifiers occur both in continuous and discrete cases as consequences of choosing specific forms of the class-conditional densities $p(\mathbf{x} | C_k)$
- Looking first at continuous input variables \mathbf{x}
- Then discussing discrete inputs

Continuous Inputs: Gaussians

- Assume Gaussian class-conditional densities with same covariance matrix Σ

$$p(\mathbf{x} | C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_k)^T \Sigma^{-1} (\mathbf{x} - \mu_k) \right\}$$

- Consider first two-class case.

– Substituting into $p(C_1 | \mathbf{x}) = \sigma \left(\ln \frac{p(\mathbf{x} | C_1)p(C_1)}{p(\mathbf{x} | C_2)p(C_2)} \right)$

– And rearranging we get $p(C_1 | \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$

- where

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2)$$

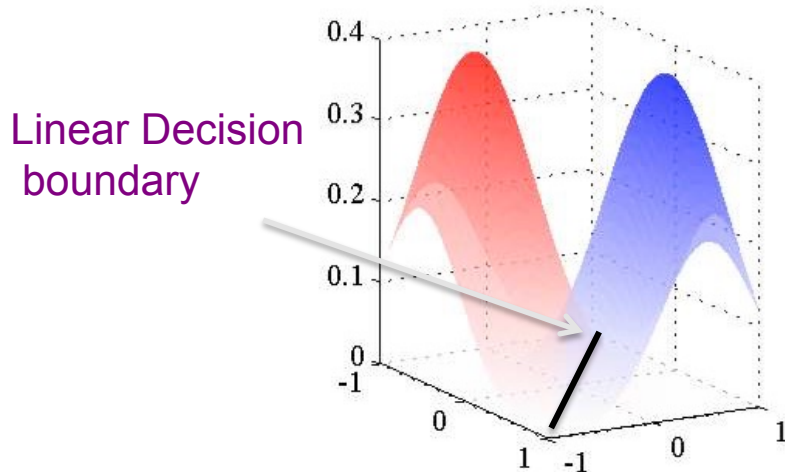
$$w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}$$

- Quadratic terms in \mathbf{x} from the exponents of the Gaussians have cancelled due to common covariance matrices
- The argument of the logistic sigmoid is a linear function of \mathbf{x}

Two Gaussian Classes

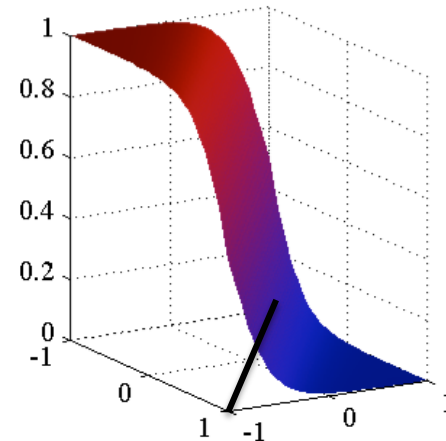
Two-dimensional input space $\mathbf{x} = (x_1, x_2)$

Class-conditional densities $p(\mathbf{x} | C_k)$



Values are positive (need not sum to 1)

Posterior $p(C_1 | \mathbf{x})$



A logistic sigmoid
of a linear function of \mathbf{x}
Red ink proportional to $p(C_1 | \mathbf{x})$
Blue ink to $p(C_2 | \mathbf{x}) = 1 - p(C_1 | \mathbf{x})$
Value 1 or 0

Continuous case with $K > 2$

$$\begin{aligned} p(C_k | \mathbf{x}) &= \frac{p(\mathbf{x} | C_k) p(C_k)}{\sum_j p(\mathbf{x} | C_j) p(C_j)} \\ &= \frac{\exp(a_k)}{\sum_j \exp(a_j)} \end{aligned}$$

- With Gaussian class conditionals

$$a_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$$

– where

$$\mathbf{w}_k = \Sigma^{-1} \mu_k$$

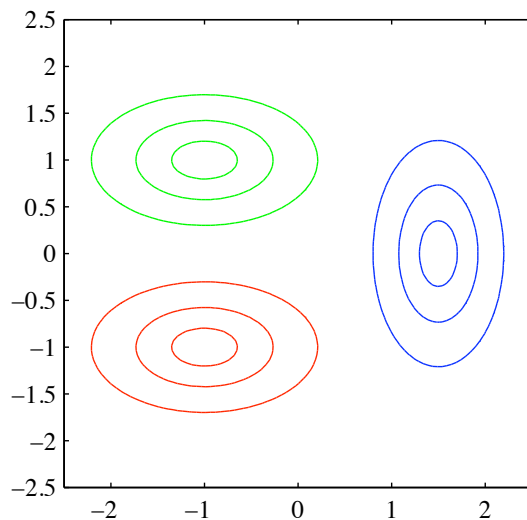
$$w_{k0} = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln p(C_k)$$

Quadratic terms
cancel thereby
leading to linearity

- If we did not assume shared covariance matrix we get a quadratic discriminant

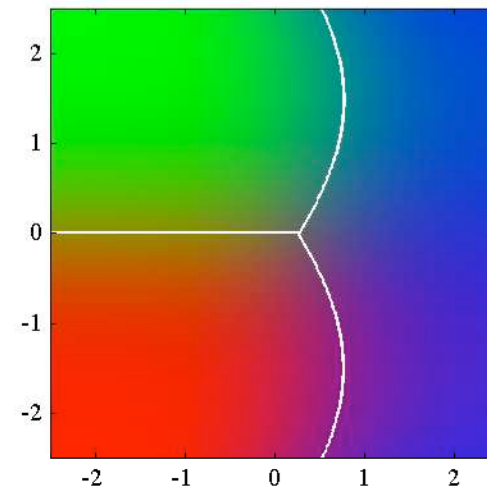
Three-class case with Gaussian models

Both Linear and Quadratic Decision boundaries



Class-conditional Densities

C_1 and C_2 have same covariance matrix



Posterior Probabilities

Between C_1 and C_2 boundary is linear,
Others are quadratic
RGB values correspond to posterior probabilities

Maximum Likelihood Solutions

- Once we have specified a parametric functional forms
 - for the class-conditional densities $p(\mathbf{x} | C_k)$
 - we can then determine the parameters together with the prior probabilities $p(C_k)$ using maximum likelihood
- This requires a data set of observations \mathbf{x} along with their class labels

M.L.E. for Gaussian Parameters

- Assuming parametric forms for $p(\mathbf{x} | C_k)$ we can determine values of parameters and priors $p(C_k)$ using maximum likelihood

Data set given $\{\mathbf{x}_n, t_n\}, n = 1, \dots, N$, $t_n = 1$ denotes class C_1 and $t_n = 0$ denotes class C_2

Let prior probabilities $p(C_1) = \pi$ $p(C_2) = 1 - \pi$

$$p(\mathbf{x}_n, C_1) = p(C_1)p(\mathbf{x}_n | C_1) = \pi \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma)$$

$$p(\mathbf{x}_n, C_2) = p(C_2)p(\mathbf{x}_n | C_2) = (1 - \pi) \mathcal{N}(\mathbf{x}_n | \mu_2, \Sigma)$$

Likelihood is given by

$$p(\mathbf{t} | \pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^N [\pi \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma)]^{t_n} [(1 - \pi) \mathcal{N}(\mathbf{x}_n | \mu_2, \Sigma)]^{1-t_n}$$

where $\mathbf{t} = (t_1, \dots, t_N)^T$

Convenient to maximize log of likelihood

Max Likelihood for Prior and Means

Estimates for prior probabilities

Log likelihood function that depend on π are $\sum_{n=1}^N \{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)\}$

MLE for p is
Fraction of points

Setting derivative to zero and rearranging $\pi = \frac{1}{N} \sum_{n=1}^N t_n = \frac{N_1}{N_1 + N_2}$ where N_1 is no fo
data points in class C_1 and N_2 in class C_2 .

Estimates for class means

Now consider maximization w.r.t. μ_1 . Pick log likelihood function depending only on μ_1

$$\sum_{n=1}^N t_n \ln \mathcal{N}(x_n | \mu_1, \Sigma) = -\frac{1}{2} \sum_{n=1}^N t_n (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) + \text{const}$$

Setting derivative to zero and solving $\mu_1 = \frac{1}{N_1} \sum_{n=1}^{N_1} t_n x_n$ ← Mean of all input vectors
 x_n assigned to class C_1

Similarly $\mu_2 = \frac{1}{N_2} \sum_{n=1}^N (1 - t_n) x_n$

Max Likelihood for Covariance Matrix

Solution for Shared Covariance Matrix

Pick out terms in log-likelihood function depending on Σ

Now maximize w.r.t. Σ

$$\begin{aligned}
 & -\frac{1}{2} \sum_{n=1}^N t_n \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^N t_n (\mathbf{x}_n - \mu_1)^T \Sigma^{-1} (\mathbf{x}_n - \mu_1) \\
 & -\frac{1}{2} \sum_{n=1}^N (1 - t_n) \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^N (1 - t_n) (\mathbf{x}_n - \mu_2)^T \Sigma^{-1} (\mathbf{x}_n - \mu_2) \\
 & = -\frac{N}{2} \ln |\Sigma| - \frac{N}{2} \text{Tr} \{ \Sigma^{-1} \mathbf{S} \}
 \end{aligned}$$

$$\mathbf{S} = \frac{N_1}{N} \mathbf{S}_1 + \frac{N_2}{N} \mathbf{S}_2$$

$$\mathbf{S}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mu_1) (\mathbf{x}_n - \mu_1)^T$$

$$\mathbf{S}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mu_2) (\mathbf{x}_n - \mu_2)^T$$

Weighted average of the two separate covariance matrices

Setting derivative to zero and solving $\Sigma = \mathbf{S}$

Discrete Features

- Assuming binary features $x_i \in \{0,1\}$
With M inputs, distribution is a table of 2^M values
- Naive Bayes assumption: independent features
Class-conditional distributions have the form

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^M \mu_{ki}^{x_i} (1 - \mu_{ki})^{1-x_i}$$

Substituting in the form needed for normalized exponential

$$\begin{aligned} a_k(\mathbf{x}) &= \ln(p(\mathbf{x} \mid C_k)p(C_k)) \\ &= \sum_{i=1}^M \left\{ x_i \ln \mu_{ki} + (1 - x_i) \ln(1 - \mu_{ki}) \right\} + \ln p(C_k) \end{aligned}$$

which is linear in \mathbf{x}

- Similar results for discrete variables which take more than 2 values

Exponential Family

- We have seen that for both Gaussian distributed and discrete inputs, the posterior class probabilities are given by generalized linear models with logistic sigmoid ($K=2$) or softmax ($K\geq 2$) activation functions
- These are particular cases of a more general result obtained by assuming that the class-conditional densities $p(\mathbf{x} | C_k)$ are members of the exponential family of distributions

Exponential Family Definition

- Class-conditionals that belong to the exponential family have the general form

$$p(\mathbf{x} \mid \lambda_k) = h(\mathbf{x})g(\lambda_k) \exp \left\{ \lambda_k^T \mathbf{u}(\mathbf{x}) \right\}$$

- Where λ_k are natural parameters of the distribution, $\mathbf{u}(\mathbf{x})$ is a function of \mathbf{x} and $g(\lambda_k)$ is a coefficient that ensures distribution is normalized
- Restricting attention to the subclass of such distributions for which $\mathbf{u}(\mathbf{x}) = \mathbf{x}$ and introducing a scaling parameter s we obtain the form

$$p(\mathbf{x} \mid \lambda_k, s) = \frac{1}{s} h\left(\frac{1}{s} \mathbf{x}\right) g(\lambda_k) \exp \left\{ \frac{1}{s} \lambda_k^T \mathbf{x} \right\}$$

- Note that each class has its own parameter vector λ_k but share a scale parameter

Exponential Family Sigmoidal form

- For the two-class problem

- Substitute expressions for the class conditional densities into $a = \ln \frac{p(\mathbf{x} | C_1)p(C_1)}{p(\mathbf{x} | C_2)p(C_2)}$ and we see that the posterior probability is given by a logistic sigmoid acting on a linear function $a(\mathbf{x})$

$$a(\mathbf{x}) = (\lambda_1 - \lambda_2)^T \mathbf{x} + \ln g(\lambda_1) - \ln g(\lambda_2) + \ln p(C_1) - \ln p(C_2)$$

- For the K -class problem

- Substituting the class-conditional density expression into $a_k = \ln \frac{p(\mathbf{x} | C_k)p(C_k)}{p(\mathbf{x} | C_1)p(C_1)}$ and we get

$$a_k(\mathbf{x}) = \lambda_k^T \mathbf{x} + \ln g(\lambda_k) + \ln p(C_k)$$

- which is again a linear function of \mathbf{x}

Summary of probabilistic linear classifiers

- Defined using
 - logistic sigmoid

$p(C_1 | \mathbf{x}) = \sigma(a)$ where a is LLR with Bayes odds

- soft-max functions

$$p(C_k | \mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

- Continuous case with shared covariance
 - we get linear functions of input \mathbf{x}
- Discrete case with independent features also results in linear functions