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 $p(\mathrm{x}|C_k)$ and priors $p(C_\mathrm{k})$
    2. Use Bayes theorem to determine posterior probabilities

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

2. Discriminative Models (One-step)
    – Directly infer posterior probabilities $p(C_k|\mathrm{x})$

- Decision Theory
    – In both cases use decision theory to assign each new x to a class

Generative Model

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

Since

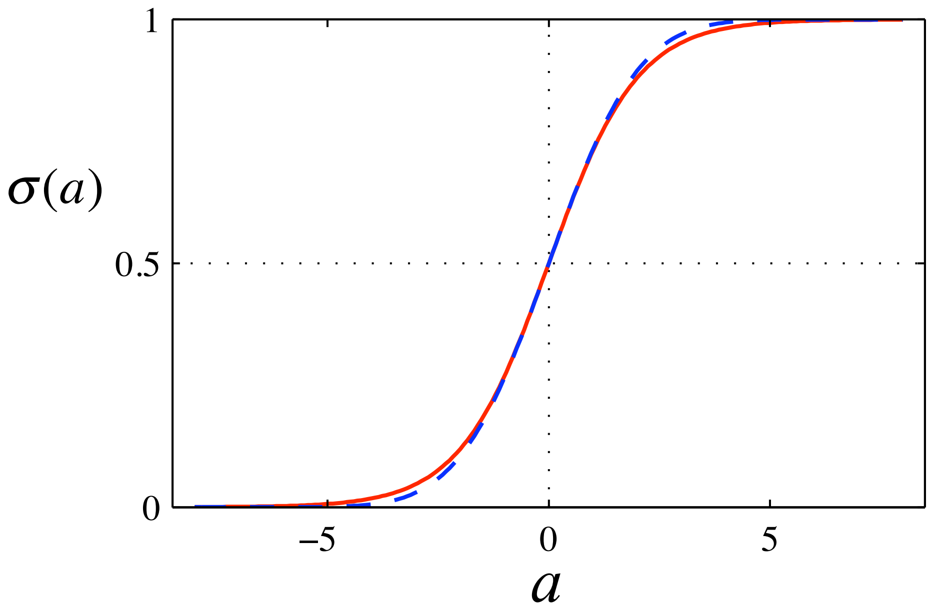
$$p(\mathrm{x}) = \sum_i p(\mathrm{x}, C_i) = \sum_i p(\mathrm{x}/C_i) p(C_i)$$

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

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

LLR with Bayes odds

Logistic Sigmoid Function



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

$$\text{Property}: \sigma(-a) = 1 - \sigma(a)$$

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

If $\sigma(a) = P(C_1 | \mathrm{x})$ then Inverse represents $\ln[p(C_1|\mathrm{x})/p(C_2|\mathrm{x})]$

Log ratio of probabilities called logit or log odds

Sigmoid: "S"-shaped or squashing function

maps real $a \ \varepsilon \ (-\infty, +\infty)$ 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 $\mathbf{x}$
- Discrete Features
- Exponential Family

Softmax: Generalization of logistic sigmoid

- For $K=2$ we have obtained logistic sigmoid
- For $K > 2$, we have its generalization

$$p(C_k \mid \mathrm{x}) = \frac{p(\mathrm{x} \mid C_k)p(C_k)}{\sum_j p(\mathrm{x} \mid C_j)p(C_j)}$$

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

If $K=2$ this reduces to earlier form

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

$$a = \ln \frac{p(\mathrm{x} \mid C_1)p(C_1)}{p(\mathrm{x} \mid C_2)p(C_2)}$$

  - Quantities a_k are defined by

$$a_k = \ln\, p(\mathrm{x} | C_k)p(C_k)$$

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

From Sigmoid to Softmax

- Binary case: we wished to produce a single no.

$$\boxed{\hat{y} = P(y = 1 \mid \boldsymbol{x})}$$

  - Since (i) this number needed to lie between 0 and 1 and (ii) because we wanted its logarithm to be well-behaved for a gradient-based optimization of log-likelihood, we chose instead to predict a number

$$\boxed{z = \log \tilde{P}(y = 1 \mid \boldsymbol{x})}$$

  - Exponentiating and normalizing, gave us a Bernoulli distribution controlled by the sigmoidal transformation of z

$$\boxed{\begin{array}{l} \log \tilde{P}(y) = yz \\ \tilde{P}(y) = \exp(yz) \end{array}} \qquad \boxed{P(y) = \frac{\exp(yz)}{\sum_{y'=0}^{1} \exp(y'z)} = \sigma((2y-1)z)}$$

- Case of n values: need to produce vector $\hat{\boldsymbol{y}}$
  - with values $\boxed{\hat{y}_i = P(y = i \mid \boldsymbol{x})}$

Softmax definition

- We need to produce a vector $\hat{\boldsymbol{y}}$ with values

$$\hat{y}_i = P(y = i \mid \boldsymbol{x})$$

  - We need elements of $\hat{\boldsymbol{y}}$ lie in $[0,1]$ and they sum to 1

- Same approach as with Bernoulli works for Multinoulli distribution

$$\boxed{z_i = \log \hat{P}(y = i \mid \boldsymbol{x})}$$

- Softmax can then exponentiate and normalize $\boldsymbol{z}$ to obtain the desired
- Softmax is given by: $\hat{\boldsymbol{y}}$

$$\boxed{\text{softmax}(\boldsymbol{z})_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}}$$

Specific forms of class-conditionals

- We next look at consequences of choosing specific forms of the class-conditional densities $p(\mathrm{x}|C_k)$
- Looking first at continuous input variables x
- Then discussing discrete inputs

Continuous Inputs

- Assume Gaussian class-conditional densities with same covariance matrix

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

- First consider two-class case

$$p(C_1 \mid \mathrm{x}) = \sigma\left(\ln \frac{p(\mathrm{x} \mid C_1) p(C_1)}{p(\mathrm{x} \mid C_2) p(C_2)} \right) = \sigma(\mathrm{w}^T \mathrm{x} + w_0)$$

  – where

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

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

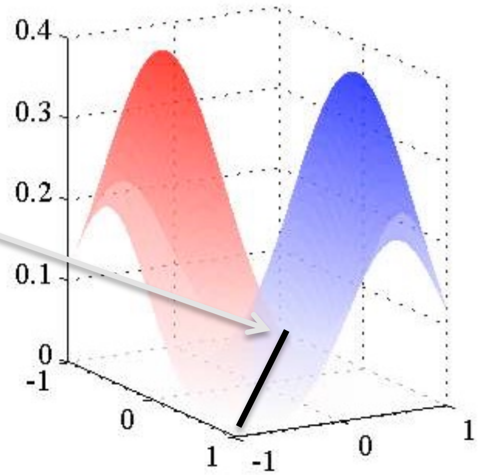
  – Quadratic terms in x cancel due to common *covariance*
  – A linear function of x in argument of logistic sigmoid

Two Gaussian Classes

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

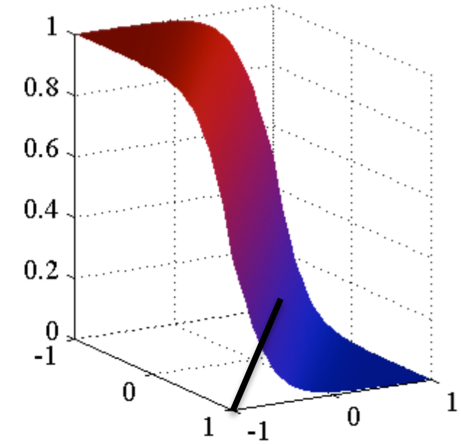
Class-conditional densities $p(\mathrm{x}|C_k)$

Linear Decision boundary



Values are positive (need not sum to 1)

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



A logistic sigmoid
of a linear function of x

Red ink proportional to $p(C_1/\mathrm{x})$
Blue ink to $p(C_2/\mathrm{x})=1\text{-}p(C_1/\mathrm{x})$
Value 1 or 0

Continuous case with $K > 2$

$$p(C_k \mid \mathrm{x}) = \frac{p(\mathrm{x} \mid C_k)p(C_k)}{\sum_j p(\mathrm{x} \mid C_j)p(C_j)}$$

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

- With Gaussian class conditionals

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

  - where

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

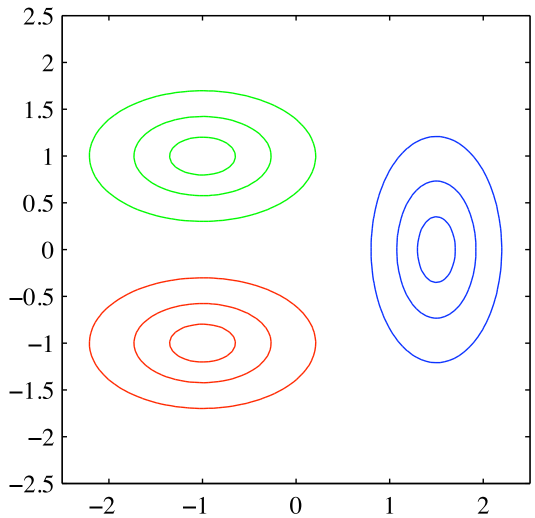
$$w_{k0} = -\frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + \ln p(C_k)$$

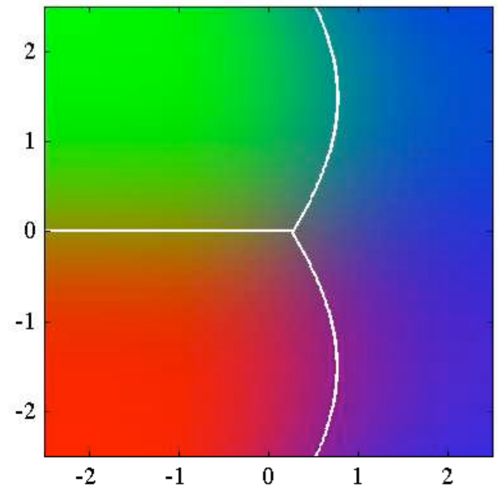
Quadratic terms cancel thereby leading to linearity

  - 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 $p(\mathrm{x}|C_k)$
  - we can then determine the parameters together with the prior probabilities $p(C_k)$ using maximum likelihood
- This requires a data set of observations x along with their class labels

M.L.E. for Gaussian Parameters

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

Data set given $\{\mathrm{x}_n, t_n\}, n = 1, \ldots, N,$ $t_n = 1$ denotes class $\mathcal{C}_1$ and $t_n = 0$ denotes class $\mathcal{C}_2$

Let prior probabilities $p(\mathcal{C}_1) = \pi$ $p(\mathcal{C}_2) = 1 - \pi$

$p(\mathrm{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathrm{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathrm{x}_n|\mu_1, \Sigma)$

$p(\mathrm{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathrm{x}_n|\mathcal{C}_2) = (1 - \pi)\mathcal{N}(\mathrm{x}_n|\mu_2, \Sigma)$

Likelihood is given by

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

where $\mathrm{t} = (t_1, .., t_N)^{\mathrm{T}}$

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 $\boxed{\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N_1+N_2}}$ 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

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

Setting derivative to zero and solving $\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n$

Mean of all input vectors x_n assigned to class C_1

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

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

$$-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\mathbf{x}_n \boldsymbol{\mu}_1)$$

$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(1-t_n)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_2)$$

$$= -\frac{N}{2}\ln|\Sigma| - \frac{N}{2}\mathrm{Tr}\left\{\Sigma^{-1}\mathbf{S}\right\}$$

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

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

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

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 D inputs, distribution is a table of 2^D values
- Naive Bayes assumption: independent features
  - Class-conditional distributions have the form

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

  - Substituting in the form needed for normalized exponential

$$a_k(\mathrm{x}) = \ln(p(\mathrm{x} \mid C_k)\, p(C_k))$$

$$= \sum_{i=1}^{D} \{x_i \ln \mu_{ki} + (1-x_i)\ln(1-\mu_{ki}\} + \ln p(C_k)$$

  - which is linear in x
- Similar results for discrete variables which take $\mathrm{M}>2$ values

Exponential Family

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

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

  - Where η are natural parameters of the distribution, $\mathrm{u}(\mathrm{x})$ is a function of x and $\mathrm{g}(\eta)$ is a coefficient that ensures distribution is normalized
  - Bernoulli, Multinomial and Gaussian belong
- For $K \geq 2$

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

  - we get $a_k(\mathrm{x}) = \lambda_k^T \mathrm{x} + \ln g(\lambda_k) + \ln p(C_k)$
  - which is linear in x

Summary of probabilistic linear classifiers

- Defined using
  - logistic sigmoid

    $$P(C_1 \mid \mathrm{x}) = \sigma(a) \quad \text{where } a \text{ is LLR with Bayes odds}$$

  - soft-max functions

    $$P(C_k \mid \mathrm{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

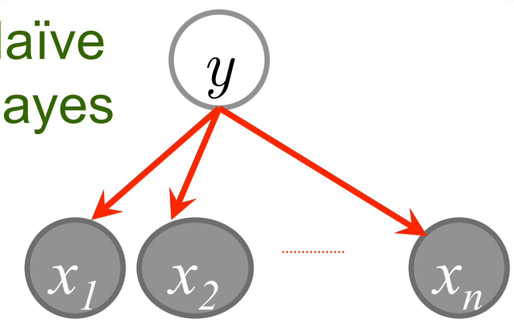
- Continuous case with shared covariance
  - we get linear functions of input x
- Discrete case with independent features also results in linear functions

Generative vs Discriminative Training

Independent variables $\mathrm{x} = \{x_1,..x_n\}$ and binary target y

1. Generative: estimate CPD parameters

Naïve Bayes



$$P(y,\mathrm{x}) = P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

From joint
get required
conditional

Low-dimensional estimation

independently estimate n x D parameters

But independence is false

For sparse data generative is better

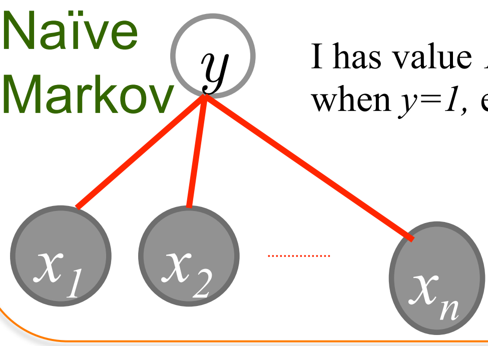
2. Discriminative: estimate CRF parameters w_i

Potential Functions (log-linear)

$\phi_i(x_i, y) = \exp\{\mathrm{w}_i x_i \,\mathrm{I}\{y=1\}\}$,

$\phi_0(y) = \exp\{\mathrm{w}_0 \,\mathrm{I}\{y=1\}\}$

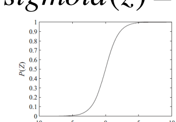
Naïve Markov



I has value 1
when $y=1$, else 0

Unnormalized $\quad \tilde{P}(y=1 \mid \mathrm{x}) = \exp\left\{\mathrm{w}_0 + \sum_{i=1}^{n} \mathrm{w}_i x_i\right\} \qquad \tilde{P}(y=0 \mid \mathrm{x}) = \exp\{0\} = 1$

Normalized $\quad P(y=1 \mid \mathrm{x}) = sigmoid\left\{\mathrm{w}_0 + \sum_{i=1}^{n} \mathrm{w}_i x_i\right\}$ where $sigmoid(z) = \frac{e^z}{1+e^z}$



Logistic Regression

Jointly optimize 12 parameters

High dimensional estimation
but correlations accounted for
Can use much richer features:
Edges, image patches sharing same pixels

multiclass

$$p(y_i \mid \phi) = y_i(\phi) = \frac{\exp(a_i)}{\sum_j \exp(a_j)}$$

where $a_j = \mathrm{w}_j^{\mathrm{T}} \phi$