Linear Classification: Probabilistic Generative Models

Karan Nathwani

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

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

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

- Decision Theory
    – In both cases use decision theory to assign each new $\boldsymbol{x}$ 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

Since

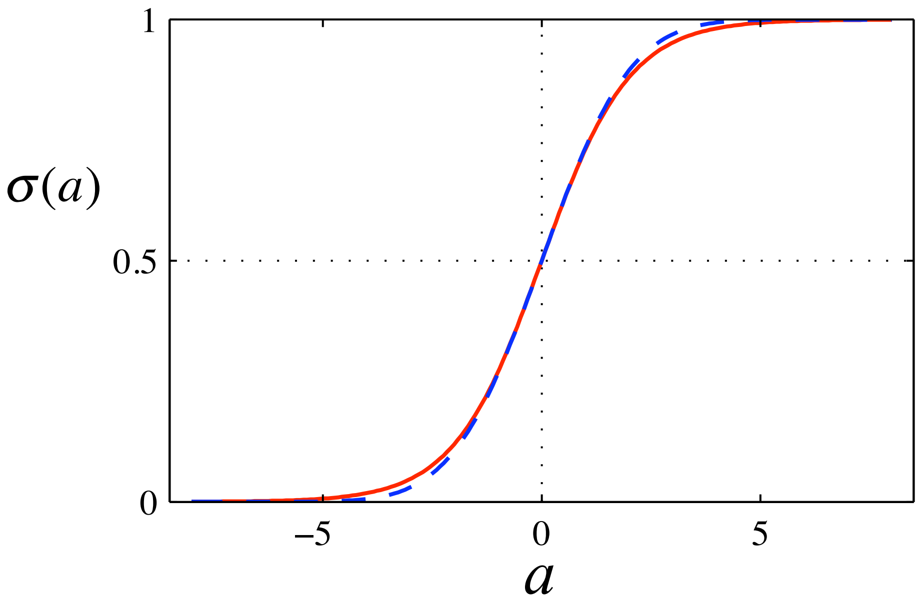
$$p(\boldsymbol{x}) = \sum_i p(\boldsymbol{x}, C_i) = \sum_i p(\boldsymbol{x}|C_i)p(C_i)$$

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

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

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 $\sigma(a) = P(C_1 \mid \boldsymbol{x})$ then inverse represents $\ln[p(C_1|\boldsymbol{x})/p(C_2|\boldsymbol{x})$

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

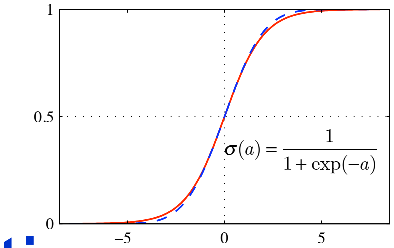
cdf of a zero-mean unit variance Gaussian

Generalizations and Special Cases

- More than 2 classes
- Gaussian Distribution of $\boldsymbol{x}$
- 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

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

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

If K=2 this reduces to a sigmoid

$p(C_1|\mathbf{x})=\exp(a_1)/[\exp(a_1)+\exp(a_2)]$

$=1/[1+\exp(a_2-a_1)]$

$=1/[1+\exp(\ln p(\mathbf{x}|C_2)p(C_2)-\ln(\mathbf{x}|C_1)p(C_1)]$

$=1/[1+p(x|C_2)p(C_2)/p(\mathbf{x}|C_1)p(C_1)]$

$=1/[1+\exp(-a)]$ where

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

  – 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 $a_k>>a_j$ 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 $p(\boldsymbol{x}|C_k)$
- Looking first at continuous input variables $\boldsymbol{x}$
- Then discussing discrete inputs

Continuous Inputs: Gaussians

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

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

- Consider first two-class case.
  - Substituting into $p(C_1 \mid \boldsymbol{x}) = \sigma\left(\ln \frac{p(\boldsymbol{x} \mid C_1)p(C_1)}{p(\boldsymbol{x} \mid C_2)p(C_2)}\right)$
  - And rearranging we get $\boxed{p(C_1 \mid \boldsymbol{x}) = \sigma(\boldsymbol{w}^T \boldsymbol{x} + w_0)}$
    - where

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

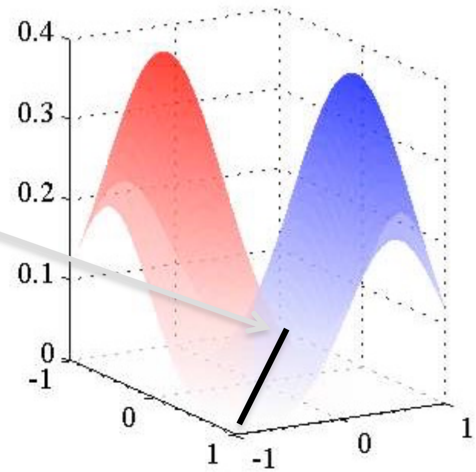
  - Quadratic terms in $\boldsymbol{x}$ from the exponents of the Gaussians have cancelled due to common covariance matrices
- The argument of the logistic sigmoid is a linear function of $\boldsymbol{x}$

Two Gaussian Classes

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

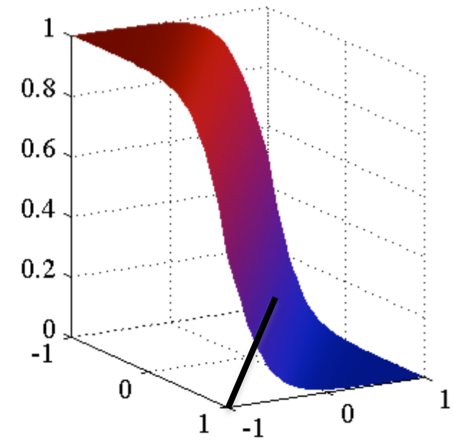
Class-conditional densities $p(\boldsymbol{x}|C_k)$

Linear Decision boundary



Values are positive (need not sum to 1)

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



A logistic sigmoid of a linear function of $\boldsymbol{x}$

Red ink proportional to $p(C_1|\boldsymbol{x})$

Blue ink to $p(C_2|\boldsymbol{x})=1-p(C_1|\boldsymbol{x})$

Value 1 or 0

Continuous case with $K > 2$

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

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

- With Gaussian class conditionals

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

  – where

$$\boldsymbol{w}_k = \Sigma^{-1} \boldsymbol{\mu}_k$$

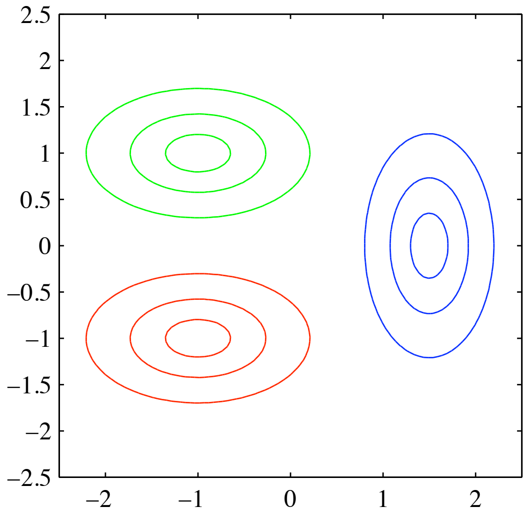
$$w_{k0} = -\frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \ln p(C_k)$$

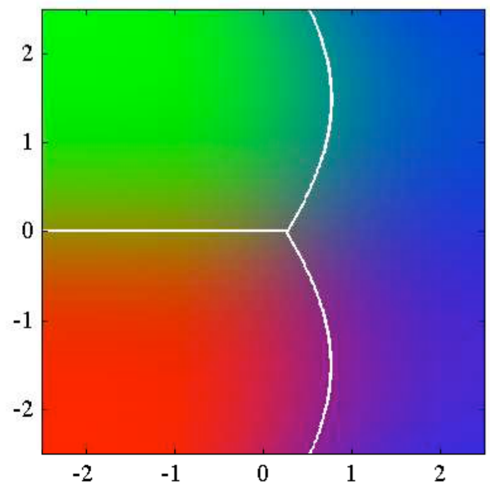
Quadratic terms cancel thereby leading to linearity

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

M.L.E. for Gaussian Parameters

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

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

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

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

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

Likelihood is given by

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) == \prod_{n=1}^{N} [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

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

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 π 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 $\mathcal{C}_1$ and N_2 in class $\mathcal{C}_2$.

Estimates for class means

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

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

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 $\mathbf{x}_n$ 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 M inputs, distribution is a table of 2^M values

- Naive Bayes assumption: independent features
  Class-conditional distributions have the form

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

Substituting in the form needed for normalized exponential

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

which is linear in $\boldsymbol{x}$

- 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 $p(\boldsymbol{x}|C_k)$ are members of the exponential family of distributions

Exponential Family Definition

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

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

  – Where λ_k are natural parameters of the distribution, $u(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ and $g\,(\lambda_k)$ is a coefficient that ensures distribution is normalized

- Restricting attention to the subclass of such distributions for which $\boldsymbol{u}(\boldsymbol{x})=\boldsymbol{x}$ and introducing a scaling parameter s we obtain the form

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

  - 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 $a = \ln \frac{p(\boldsymbol{x} \mid C_1)p(C_1)}{p(\boldsymbol{x} \mid C_2)p(C_2)}$ and we see that the posterior probability is given by a logistic sigmoid acting on a linear function $a(\boldsymbol{x})$

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

- For the K-class problem
  - Substituting the class-conditional density expression into $a_k = \ln p(\boldsymbol{x}|C_k)p(C_k)$ and we get

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

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

Summary of probabilistic linear classifiers

- Defined using
  - logistic sigmoid

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

  - soft-max functions

$$p(C_k \mid \boldsymbol{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

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