Probabilistic generative models

Linear models for classification

Francesco Corona

Probabilistic generative models

Probabilistic discriminative models

Models with linear decision boundaries arise from assumptions about the data

In a generative approach to classification, we first model the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ and the class priors $p(\mathcal{C}_k)$, and then

- we compute posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ through Bayes' theorem

Probabilistic discriminative models (cont.)

For the two-class problem, the posterior probability of class $\mathcal{C}_1$ is

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{\underbrace{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}_{p(\mathbf{x})=\sum_k p(\mathbf{x},\mathcal{C}_k)=\sum_k p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}} = \frac{1}{1+\exp(-a)} = \sigma(a) \qquad (1)$$

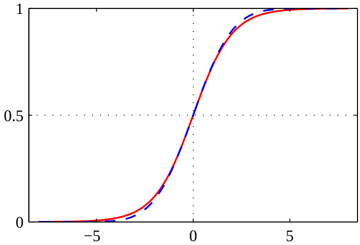
where we defined

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \qquad (2)$$

$\sigma(a)$ is the **logistic sigmoid function** (plotted in red)

$$\sigma(a) = \frac{1}{1+\exp(-a)} \qquad (3)$$

or **squashing function**, because it maps $\mathbb{R}$ onto a finite interval



Probabilistic discriminative models (cont.)

The logistic sigmoid satisfies the following symmetry property

$$\sigma(-a) = 1 - \sigma(a) \tag{4}$$

The inverse of the logistic sigmoid is known as **logit function**

$$a = \ln\left(\frac{\sigma}{1-\sigma}\right) \tag{5}$$

It reflects the log of the ratio of probabilities for two classes

$$\ln\left(p(\mathcal{C}_1|\mathbf{x})/p(\mathcal{C}_2|\mathbf{x})\right)$$

Probabilistic discriminative models (cont.)

$$
\begin{aligned}
p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \frac{1}{1 + \exp\Big(- \underbrace{\ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}}_{a}\Big)} \\
&= \sigma\Big(\underbrace{\ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}}_{a}\Big)
\end{aligned}
$$

We have written the posterior probabilities in an equivalent form that will have significance when $a(\mathbf{x})$ is a linear function of $\mathbf{x}$

- Here, the posterior probability is governed by a generalised linear model

Probabilistic discriminative models (cont.)

For the case $K > 2$ classes, we have

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{j=1}^{K} p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp{(a_k)}}{\sum_{j=1}^{K} \exp{(a_j)}} \tag{6}$$

known as **normalised exponential**[1]

We have defined the quantity a_k as

$$a_k = \ln \Big(p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k) \Big) \tag{7}$$

If $a_k >> a_j$, for all $j \neq k$, then $\begin{cases} p(\mathcal{C}_k|\mathbf{x}) & \simeq 1 \\ p(\mathcal{C}_j|\mathbf{x}) & \simeq 0 \end{cases}$

We analyse the consequences of choosing the form of class-conditional densities

[1] It is a generalisation of the logistic sigmoid and it is also known as the softmax function

Outline

Probabilistic generative models
- Continuous inputs
- Maximum likelihood solution

Continuous inputs
Probabilistic generative models

Continuous inputs

Let us assume that the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ are Gaussian

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\Big(- \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\Big) \qquad (8)$$

and, we want to explore the form of the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$

The Gaussians have different means $\boldsymbol{\mu}_k$ but share the same covariance matrix $\boldsymbol{\Sigma}$

Continuous inputs (cont.)

With 2 classes, $p(\mathcal{C}_1|\mathbf{x}) = \dfrac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \dfrac{1}{1+\exp(-a)} = \sigma(a)$

and $a = \ln \dfrac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$, we have

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0) \tag{9}$$

where

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{10}$$

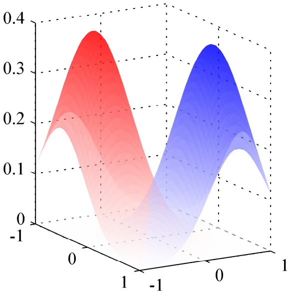
$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \tag{11}$$

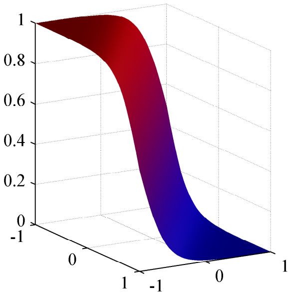
The quadratic terms in $\mathbf{x}$ from the exponents of the Gaussian densities have cancelled (due to the assumption of common covariance matrices) leading to

- a linear function of $\mathbf{x}$ in the argument of the logistic sigmoid

Continuous inputs (cont.)

The left-hand plot shows the class-conditional densities for two classes over $2D$





The posterior probability $p(\mathcal{C}_1|\mathbf{x})$ is a logistic sigmoid of a linear function of $\mathbf{x}$

The surface in the right-hand plot is coloured using a proportion of red given by $p(\mathcal{C}_1|\mathbf{x})$ and a proportion of blue given by $p(\mathcal{C}_2|\mathbf{x}) = 1 - p(\mathcal{C}_1|\mathbf{x})$

Continuous inputs (cont.)

Decision boundaries are surfaces with constant posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$

- Linear functions of $\mathbf{x}$
- Linear in input space

Prior probabilities $p(\mathcal{C}_k)$ enter only through the bias parameter w_0 so changes in priors have the effect of making parallel shifts of the decision boundary

- more generally of the parallel contours of constant posterior probability

Continuous inputs (cont.)

For the K-class case, using $p(\mathcal{C}_k|\mathbf{x}) = \dfrac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{j=1}^K p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \dfrac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)}$ and $a_k = \ln(p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k))$, we have

$$a_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0} \tag{12}$$

$$\mathbf{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \tag{13}$$

$$w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k) \tag{14}$$

The $a_k(\mathbf{x})$ are again linear functions of $\mathbf{x}$ as a consequence of the cancellation of the quadratic terms due to the shared covariances

The resulting decision boundaries (minimum misclassification rate) occur when two of the posterior probabilities (the two largest) are equal, and so they are defined by linear functions of $\mathbf{x}$

- again we have a generalised linear model

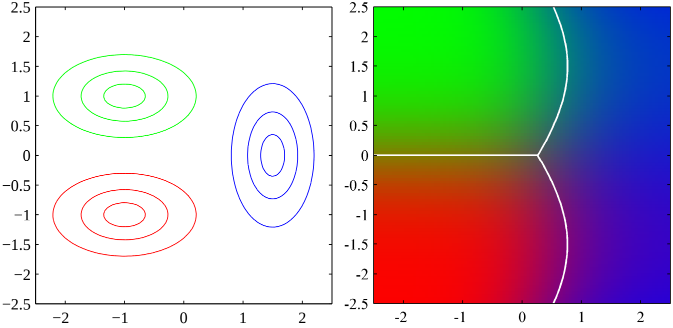
Continuous inputs (cont.)

If we relax the assumption of a shared covariance matrix and allow each class-conditional density $p(\mathbf{x}|\mathcal{C}_k)$ to have its own covariance matrix $\boldsymbol{\Sigma}_k$, then the earlier cancellations will no longer occur, and we will obtain quadratic functions of $\mathbf{x}$, giving rise to a **quadratic discriminant**

Continuous inputs (cont.)

Class-conditional densities for three classes each having a Gaussian distribution

- red and green classes have the same covariance matrix



The corresponding posterior probabilities and the decision boundaries

- Linear boundary between red and green classes, same covariance matrix
- Quadratic boundaries between other pairs, different covariance matrix

Maximum likelihood solution

Probabilistic generative models

Maximum likelihood solution

Once we specified a parametric functional form for class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$, we can determine parameters and prior class probabilities $p(\mathcal{C}_k)$

- Maximum likelihood

This requires data comprising observations of $\mathbf{x}$ and corresponding class labels

Maximum likelihood solution (cont.)

Consider first the two-class case, each having a Gaussian density with shared covariance matrix $\boldsymbol{\Sigma}$, and suppose we have data $\{\mathbf{x}_n, t_n\}_{n=1}^N$

$$\begin{cases} t_n = 1, & \text{for } \mathcal{C}_1 \text{ with prior probability } p(\mathcal{C}_1) = \pi \\ t_n = 0, & \text{for } \mathcal{C}_2 \text{ with prior probability } p(\mathcal{C}_2) = 1 - \pi \end{cases}$$

For a data point $\mathbf{x}_n$ from class $\mathcal{C}_1$ ($\mathcal{C}_2$), we have $t_n = 1$ ($t_n = 0$) and thus

$$\begin{aligned} p(\mathbf{x}_n, \mathcal{C}_1) &= p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ p(\mathbf{x}_n, \mathcal{C}_2) &= p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \end{aligned}$$

For $\mathtt{t} = (t_1, \ldots, t_n)^T$, the likelihood function is given by

$$p(\mathtt{t}, \mathtt{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \Big(\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\Big)^{t_n} \Big((1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\Big)^{1-t_n} \quad (15)$$

Maximum likelihood solution (cont.)

As usual, we maximise the log of the likelihood function

$$\sum_{n=1}^{N} \underbrace{t_n \ln(\pi) + (1-t_n)\ln(1-\pi)}_{\pi} + \underbrace{\underbrace{t_n \ln\big(\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\big)}_{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}} + \underbrace{(1-t_n)\ln\big(\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\big)}_{\boldsymbol{\mu}_2, \boldsymbol{\Sigma}}}_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}}$$

Maximum likelihood solution (cont.)

Consider first maximisation with respect to π, where the terms on π are

$$\sum_{n=1}^{N} \Big(t_n \ln(\pi) + (1 - t_n) \ln(1 - \pi) \Big) \tag{16}$$

Setting the derivative wrt π to zero and rearranging

$$\pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} \tag{17}$$

The maximum likelihood estimate for π is the fraction of points in $\mathcal{C}_1$

Maximum likelihood solution (cont.)

Now consider maximisation with respect to $\boldsymbol{\mu}_1$, where the terms on $\boldsymbol{\mu}_1$ are

$$\sum_{n=1}^{N} t_n \ln \Big(\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \Big) = -\frac{1}{2} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) + \text{const} \quad (18)$$

Setting the derivative wrt $\boldsymbol{\mu}_1$ to zero and rearranging

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n \quad (19)$$

The maximum likelihood estimate of $\boldsymbol{\mu}_1$ is the mean of inputs $\mathbf{x}_n$ in class $\mathcal{C}_1$

$$\boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} t_n \mathbf{x}_n \quad (20)$$

Maximum likelihood solution (cont.)

Lastly consider maximisation with respect to $\mathbf{\Sigma}$, where the terms on $\mathbf{\Sigma}$ are

$$
\begin{aligned}
-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n(\mathbf{x}_n - \boldsymbol{\mu}_1)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1) \\
-\frac{1}{2}\sum_{n=1}^{N} (1 - t_n) \ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (1 - t_n)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_2) \\
= -\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{N}{2}\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) \qquad (21)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{S} &= \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 && (22) \\
\mathbf{S}_1 &= \frac{1}{N_1}\sum_{n\in\mathcal{C}_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T && (23) \\
\mathbf{S}_2 &= \frac{1}{N_2}\sum_{n\in\mathcal{C}_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T && (24)
\end{aligned}
$$

Maximum likelihood solution (cont.)

$$\boldsymbol{\Sigma} = \mathbf{S} = \frac{N_1}{N}\frac{1}{N_1}\sum_{n \in \mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T + \frac{N_2}{N}\frac{1}{N_2}\sum_{n \in \mathcal{C}_2}(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T$$

It is the average of the covariance matrices associated with each class separately