Machine learning

Probabilistic classification - generative models

Corso di Laurea Magistrale in Informatica

Università di Roma Tor Vergata

Prof. Giorgio Gambosi

a.a. 2023-2024

Generative models

- Classes are modeled by suitable conditional distributions $p(\mathbf{x}|C_k)$ (language models in the previous case): it is possible to sample from such distributions to generate random documents statistically equivalent to the documents in the collection used to derive the model.
- Bayes' rule allows to derive $p(C_k|\mathbf{x})$ given such models (and the prior distributions $p(C_k)$ of classes)
- We may derive the parameters of $p(\mathbf{x}|C_k)$ and $p(C_k)$ from the dataset, for example through maximum likelihood estimation
- Classification is performed by comparing $p(C_k|\mathbf{x})$ for all classes

Deriving posterior probabilities

- Let us consider the binary classification case and observe that

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)} = \frac{1}{1 + \frac{p(\mathbf{x}|C_2)p(C_2)}{p(\mathbf{x}|C_1)p(C_1)}}$$

- Let us define

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

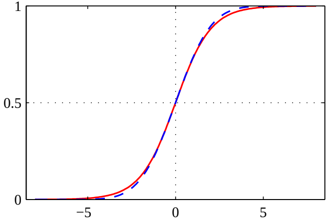
  that is, a is the log of the ratio between the posterior probabilities (log odds)

- We obtain that

$$p(C_1|\mathbf{x}) = \frac{1}{1 + e^{-a}} = \sigma(a) \qquad p(C_2|\mathbf{x}) = 1 - \frac{1}{1 + e^{-a}} = \frac{1}{1 + e^{a}}$$

- $\sigma(x)$ is the logistic function or (sigmoid)

Sigmoid



Useful properties of the sigmoid

- $\sigma(-x) = 1 - \sigma(x)$
- $\dfrac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$

Deriving posterior probabilities

- In the case $K > 2$, the general formula holds

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

- Let us define, for each $k = 1, \ldots, K$

$$a_k(\mathbf{x}) = \log(p(\mathbf{x}|C_k)p(C_k)) = \log p(\mathbf{x}|C_k) + \log p(C_k)$$

- Then, we may write

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

- $s(\mathbf{x})$ is the softmax function (or normalized exponential) and it can be seen as an extension of the sigmoid to the case $K > 2$ and as a smoothed version of the maximum

Gaussian discriminant analysis

In Gaussian discriminant analysis (GDA) all class conditional distributions $p(\mathbf{x}|C_k)$ are assumed gaussians. This implies that the corresponding posterior distributions $p(C_k|\mathbf{x})$ can be easily derived.

Hypothesis

All distributions $p(\mathbf{x}|C_k)$ have same covariance matrix $\mathbf{\Sigma}$, of size $D \times D$. Then,

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

Binary case

If $K = 2$,

$$p(C_1|\mathbf{x}) = \sigma(a(\mathbf{x}))$$

where

$$\begin{aligned} a(\mathbf{x}) &= \log \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} \\ &= \frac{1}{2}(\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\mathbf{x}) - \frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\mathbf{x}) + \log \frac{p(C_1)}{p(C_2)} \end{aligned}$$

Binary case

Observe that the results of all products involving $\mathbf{\Sigma}^{-1}$ are scalar, hence, in particular

$$\mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 = \boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\mathbf{x}$$
$$\mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\mathbf{x}$$

Then,

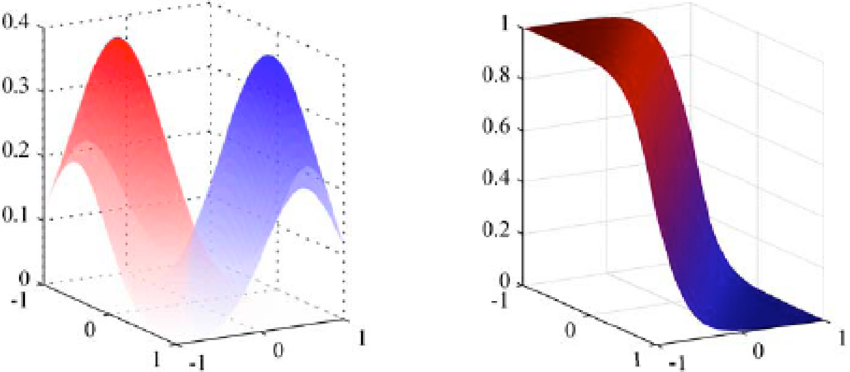
$$a(\mathbf{x}) = \frac{1}{2}(\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1) + (\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1} - \boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1})\mathbf{x} + \log\frac{p(C_1)}{p(C_2)} = \mathbf{w}^T\mathbf{x} + w_0$$

with

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$
$$w_0 = \frac{1}{2}(\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1) + \log\frac{p(C_1)}{p(C_2)}$$

$p(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$ is computed by applying a non-linear function to a linear combination of the features (generalized linear model)

Example



Left, the class conditional distributions $p(\mathbf{x}|C_1), p(\mathbf{x}|C_2)$, gaussians with $D = 2$. Right the posterior distribution of C_1, $p(C_1|\mathbf{x})$ with sigmoidal slope.

Discriminant function

The discriminant function can be obtained by the condition $p(C_1|\mathbf{x}) = p(C_2|\mathbf{x})$, that is, $\sigma(a(\mathbf{x})) = \sigma(-a(\mathbf{x}))$.
This is equivalent to $a(\mathbf{x}) = -a(\mathbf{x})$ and to $a(\mathbf{x}) = 0$. As a consequence, it results

$$\mathbf{w}^T\mathbf{x} + w_0 = 0$$

or

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) + \log\frac{p(C_2)}{p(C_1)} = 0$$

Simple case: $\boldsymbol{\Sigma} = \lambda\mathbf{I}$ (that is, $\sigma_{ii} = \lambda$ for $i = 1, \ldots, d$). In this case, the discriminant function is

$$2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathbf{x} + ||\boldsymbol{\mu}_1||^2 - ||\boldsymbol{\mu}_2||^2 + 2\lambda\log\frac{p(C_2)}{p(C_1)} = 0$$

Multiple classes

In this case, we refer to the softmax function:

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

where $a_k(\mathbf{x}) = \log(p(\mathbf{x}|C_k)p(C_k))$.
By the above considerations, it easily turns out that

$$a_k(\mathbf{x}) = \frac{1}{2}\left(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\right) + \log p(C_k) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| = \mathbf{w}_k^T\mathbf{x} + w_{0k}$$

Again, $p(C_k|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$ is computed by applying a non-linear function to a linear combination of the features (generalized linear model)

Multiple classes

Decision boundaries corresponding to the case when there are two classes C_j, C_k such that the corresponding posterior probabilities are equal, and larger than the probability of any other class. That is,

$$p(C_k|\mathbf{x}) = p(C_j|\mathbf{x}) \qquad\qquad p(C_i|\mathbf{x}) < p(C_k|\mathbf{x}) \quad i \neq j, k$$

hence

$$e^{a_k(\mathbf{x})} = e^{a_j(\mathbf{x})} \qquad\qquad e^{a_i(\mathbf{x})} < e^{a^k(\mathbf{x})} \quad i \neq j, k$$

that is,

$$a_k(\mathbf{x}) = a_j(\mathbf{x}) \qquad\qquad a_i(\mathbf{x}) < a^k(\mathbf{x}) \quad i \neq j, k$$

As shown, this implies that boundaries are linear.

General covariance matrices, binary case

The class conditional distributions $p(\mathbf{x}|C_k)$ are gaussians with different covariance matrices

$$
\begin{aligned}
a(\mathbf{x}) &= \log \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} \\
&= \frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - (\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right) + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \log\frac{p(C_1)}{p(C_2)}
\end{aligned}
$$

General covariance matrices, binary case

By applying the same considerations, the decision boundary turns out to be

$$\left((\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right)+\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}+2\log\frac{p(C_1)}{p(C_2)}=0$$

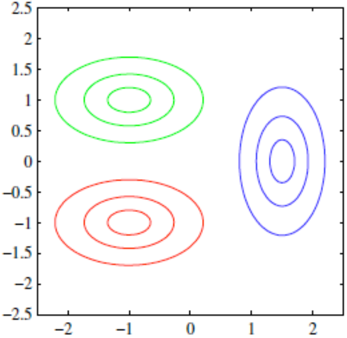
Classes are separated by a (at most) quadratic surface.

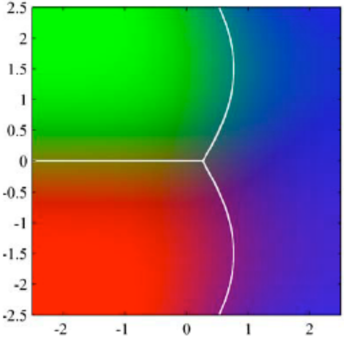
General covariance, multiple classe

It can be proved that boundary surfaces are at most quadratic.

Example

Left: 3 classes, modeled by gaussians with different covariance matrices.
Right: posterior distribution of classes, with boundary surfaces.





GDA and maximum likelihood

The class conditional distributions $p(\mathbf{x}|C_k)$ can be derived from the training set by maximum likelihood estimation.

For the sake of simplicity, assume $K = 2$ and both classes share the same $\mathbf{\Sigma}$.

It is then necessary to estimate $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, $\mathbf{\Sigma}$, and $\pi = p(C_1)$ (clearly, $p(C_2) = 1 - \pi$).

GDA and maximum likelihood

Training set $\mathcal{T}$: includes n elements $(\mathbf{x}_i, t_i)$, with

$$t_i = \begin{cases} 0 & \text{se } \mathbf{x}_i \in C_2 \\ 1 & \text{se } \mathbf{x}_i \in C_1 \end{cases}$$

If $\mathbf{x} \in C_1$, then $p(\mathbf{x}, C_1) = p(\mathbf{x}|C_1)p(C_1) = \pi \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
If $\mathbf{x} \in C_2$, $p(\mathbf{x}, C_2) = p(\mathbf{x}|C_2)p(C_2) = (1 - \pi) \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$

The likelihood of the training set $\mathcal{T}$ is

$$L(\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}|\mathcal{T}) = \prod_{i=1}^{n} (\pi \cdot \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{t_i}((1 - \pi) \cdot \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}))^{1-t_i}$$

GDA and maximum likelihood

The corresponding log likelihood is

$$l(\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}|\mathcal{T}) = \sum_{i=1}^{n} (t_i \log \pi + t_i \log(\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))) +$$

$$+ \sum_{i=1}^{n} ((1-t_i)\log(1-\pi) + (1-t_i)\log(\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})))$$

Its derivative wrt π is

$$\frac{\partial l}{\partial \pi} = \frac{\partial}{\partial \pi}\sum_{i=1}^{n}(t_i \log \pi + (1-t_i)\log(1-\pi)) = \sum_{i=1}^{n}\left(\frac{t_i}{\pi} - \frac{(1-t_i)}{1-\pi}\right) = \frac{n_1}{\pi} - \frac{n_2}{1-\pi}$$

which is equal to 0 for

$$\pi = \frac{n_1}{n}$$

GDA and maximum likelihood

The maximum wrt $\boldsymbol{\mu}_1$ (and $\boldsymbol{\mu}_2$) is obtained by computing the gradient

$$\frac{\partial l}{\partial \boldsymbol{\mu}_1} = \frac{\partial}{\partial \boldsymbol{\mu}_1}\sum_{i=1}^{n} t_i \log(\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})) = \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n} t_i(\mathbf{x}_i - \boldsymbol{\mu}_1)$$

As a consequence, we have $\frac{\partial l}{\partial \boldsymbol{\mu}_1} = 0$ for

$$\sum_{i=1}^{n} t_i \mathbf{x}_i = \sum_{i=1}^{n} t_i \boldsymbol{\mu}_1$$

hence, for

$$\boldsymbol{\mu}_1 = \frac{1}{n_1}\sum_{\mathbf{x}_i \in C_1} \mathbf{x}_i$$

Similarly, $\frac{\partial l}{\partial \boldsymbol{\mu}_2} = 0$ for

$$\boldsymbol{\mu}_2 = \frac{1}{n_2}\sum_{\mathbf{x}_i \in C_2} \mathbf{x}_i$$

GDA and maximum likelihood

Maximizing the log-likelihood wrt $\mathbf{\Sigma}$ provides

$$\mathbf{\Sigma} = \frac{n_1}{n}\mathbf{S}_1 + \frac{n_2}{n}\mathbf{S}_2$$

where

$$\mathbf{S}_1 = \frac{1}{n_1}\sum_{\mathbf{x}_i \in C_1} (\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^T$$

$$\mathbf{S}_2 = \frac{1}{n_2}\sum_{\mathbf{x}_i \in C_2} (\mathbf{x}_i - \boldsymbol{\mu}_2)(\mathbf{x}_i - \boldsymbol{\mu}_2)^T$$

GDA: DISCRETE FEATURES

- In the case of d discrete (for example, binary) features we may apply the Naive Bayes hypothesis (independence of features, given the class)
- Then, we may assume that, for any class C_k, the value of the i-th feature is sampled from a Bernoulli distribution of parameter p_{ki}; by the conditional independence hypothesis, it results into

$$p(\mathbf{x}|C_k) = \prod_{i=1}^{d} p_{ki}^{x_i}(1-p_{ki})^{1-x_i}$$

  where $p_{ki} = p(x_i = 1|C_k)$ could be estimated by ML, as in the case of language models
- Functions $a_k(\mathbf{x})$ can then be defined as:

$$a_k(\mathbf{x}) = \log(p(\mathbf{x}|C_k)p(C_k)) = \sum_{i=1}^{D}(x_i \log p_{ki} + (1-x_i)\log(1-p_{ki})) + \log p(C_k)$$

  These are still linear functions on $\mathbf{x}$.
- The same considerations can be done in the case of non binary features, where, for any class C_k, we may assume the value of the i-th feature is sampled from a distribution on a suitable domain (e.g. Poisson in the case of count data)