Probabilistic classification

Course of Machine Learning
Master Degree in Computer Science
University of Rome ``Tor Vergata''

Giorgio Gambosi

a.a. 2017-2018

Probabilistic generative models

Linear classifiers derive form simple hypotheses on posterior $p(\mathbf{x}|C_k)$ and prior $p(C_k)$ distribution of classes

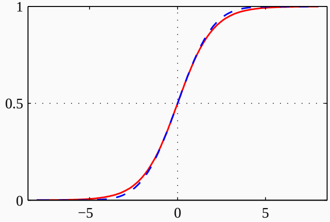
Binary case:

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

where

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

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



Sigmoid

Properties of the sigmoid

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

The inverse function of the sigmoid is the logit function

$$a = \log \frac{\sigma}{1 - \sigma}$$

As seen above, in our framework a is the log of the ratio between the posterior probabilities (log odds)

Softmax

The extension of the sigmoid to the case $K > 2$ is the softmax function (or normalized exponential)

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

where

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

Smoothed version of the maximum: if $a_k \gg a_j$ for all $j \neq k$, then $s(a_k) \simeq 1$ and $s(a_j) \simeq 0$ for all $j \neq k$

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,

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

If $K = 2$,

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

where

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

Binary case

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

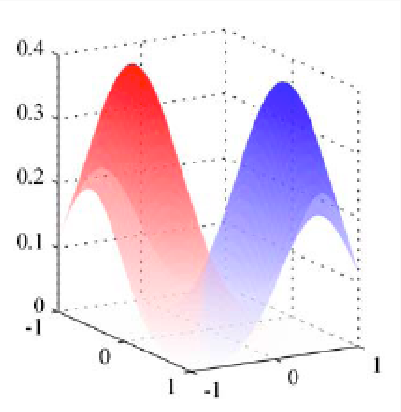
$$\mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 = \boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\mathbf{x}$$
$$\mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\mathbf{x}$$

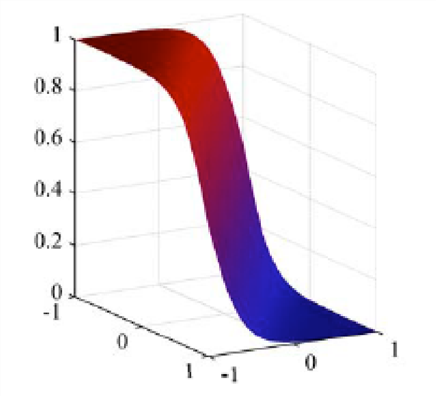
Then,

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

with

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





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 $\sigma(a(\mathbf{x})) = \sigma(-a(\mathbf{x}))$, which is equivalent to $a(\mathbf{x}) = -a(\mathbf{x})$ and to $a(\mathbf{x}) = 0$. As a consequence, it results

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

that is

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

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

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

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,

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

As shown above, this implies that boundaries are linear. In particular, $a_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{0k}$ with

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

and

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

General covariance matrices, binary case

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

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

General covariance matrices, binary case

The decision boundary is now defined by

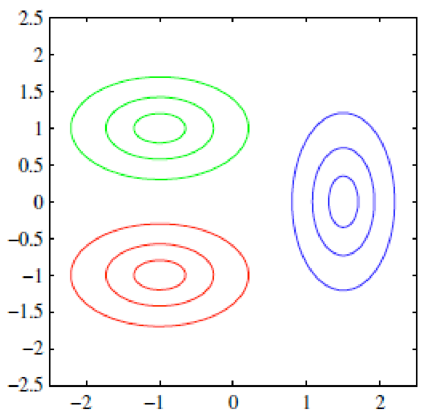
$$\Big((\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\Big)+\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}+2\log\frac{p(C_1)}{p(C_2)}=0$$

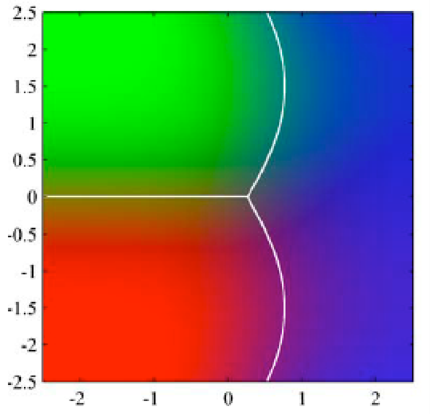
Classes are separated by a (at most) quadratic surface.

Example for the general case

Left: 3 classes, modeled by gaussians with different covariance matrices.

Right: posterior distribution of classes, with boundary surfaces.





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 $\boldsymbol{\Sigma}$.

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

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

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

The corresponding log likelihood is

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

Its derivative wrt π is

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

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

$$\frac{\partial l}{\partial \boldsymbol{\mu}_1} = \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{i=1}^{n} t_i \log(\mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})) = -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{i=1}^{n} t_i (\mathbf{x}_i - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1)$$

Let $\boldsymbol{\xi}_i = (\mathbf{x}_i - \boldsymbol{\mu}_1)$, then, by the chain rule of derivatives,

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\mu}_1} &= -\frac{1}{2} \sum_{i=1}^{n} t_i \frac{\partial}{\partial \boldsymbol{\mu}_1} \left(\boldsymbol{\xi}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi}_i \right) = -\frac{1}{2} \sum_{i=1}^{n} t_i \frac{\partial \boldsymbol{\xi}_i}{\partial \boldsymbol{\mu}_1} \frac{\partial}{\partial \boldsymbol{\xi}_i} \left(\boldsymbol{\xi}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi}_i \right) \\ &= \frac{1}{2} \sum_{i=1}^{n} t_i \left(\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^T \right) \boldsymbol{\xi}_i = \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} t_i (\mathbf{x}_i - \boldsymbol{\mu}_1)\end{aligned}$$

since in general

$$\frac{\partial}{\partial \mathbf{a}} \left(\mathbf{a}^T \mathbf{A} \mathbf{a} \right) = (\mathbf{A} + \mathbf{A}^T) \mathbf{a}$$

and $\boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Sigma}^{-1})^T$ by the symmetry of the covariance matrix.

As a consequence, we have $\dfrac{\partial l}{\partial \boldsymbol{\mu}_1} = 0$ 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, $\dfrac{\partial l}{\partial \boldsymbol{\mu}_2} = 0$ for

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

To maximize the log-likelihood wrt $\mathbf{\Sigma}$, derive the corresponding gradient

$$
\begin{aligned}
\frac{\partial l}{\partial \mathbf{\Sigma}} &= \sum_{i=1}^{n} t_i \frac{\partial}{\partial \mathbf{\Sigma}} \log(\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \mathbf{\Sigma})) + \sum_{i=1}^{n}(1-t_i)\frac{\partial}{\partial \mathbf{\Sigma}} \log(\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2, \mathbf{\Sigma})) \\
&= \sum_{i=1}^{n} t_i \frac{\partial}{\partial \mathbf{\Sigma}} \log|\mathbf{\Sigma}|^{-\frac{1}{2}} + \frac{\partial}{\partial \mathbf{\Sigma}} \Big((\mathbf{x}_i - \boldsymbol{\mu}_1)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1)\Big) \\
&+ \sum_{i=1}^{n}(1-t_i)\frac{\partial}{\partial \mathbf{\Sigma}} \log|\mathbf{\Sigma}|^{-\frac{1}{2}} + \frac{\partial}{\partial \mathbf{\Sigma}} \Big((\mathbf{x}_i - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_2)\Big) \\
&= -\frac{n}{2}\frac{\partial}{\partial \mathbf{\Sigma}} \log|\mathbf{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}} \sum_{\mathbf{x}_i \in C_1} \Big((\mathbf{x}_i - \boldsymbol{\mu}_1)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1)\Big) \\
&\quad - \frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}} \sum_{\mathbf{x}_i \in C_2} \Big((\mathbf{x}_i - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_2)\Big)
\end{aligned}
$$

Observe now that $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is a scalar, hence $\mathbf{x}^T\mathbf{A}\mathbf{x} = tr(\mathbf{x}^T\mathbf{A}\mathbf{x})$; moreover, in general

$$\mathrm{tr}\left(\mathbf{ABC}\right) = \mathrm{tr}\left(\mathbf{CAB}\right) = \mathrm{tr}\left(\mathbf{BCA}\right)$$

As a consequence, $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathrm{tr}\left(\mathbf{A}\mathbf{x}\mathbf{x}^T\right)$ and

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}} = -\frac{n}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{\mathbf{x}_i \in C_1}\mathrm{tr}\left((\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}\right)$$
$$-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{\mathbf{x}_i \in C_2}\mathrm{tr}\left((\mathbf{x}_i - \boldsymbol{\mu}_2)(\mathbf{x}_i - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\right)$$

Let us now define the following matrices

$$\mathbf{S}_1 = \frac{1}{n_1}\sum_{\mathbf{x}_i \in C_1}(\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^T$$
$$\mathbf{S}_2 = \frac{1}{n_2}\sum_{\mathbf{x}_i \in C_2}(\mathbf{x}_i - \boldsymbol{\mu}_2)(\mathbf{x}_i - \boldsymbol{\mu}_2)^T$$

and let

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

GDA and maximum likelihood

By applying these definitions, we obtain

$$\begin{aligned}
\frac{\partial l}{\partial \mathbf{\Sigma}} &= -\frac{n}{2}\frac{\partial}{\partial \mathbf{\Sigma}} \log|\mathbf{\Sigma}| - \frac{n}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\text{tr}\left(\mathbf{S}\mathbf{\Sigma}^{-1}\right) \\
&= -\frac{n}{2}\left(\mathbf{\Sigma}^{-1}\right)^T - \frac{n}{2}\frac{\partial \mathbf{\Sigma}^{-1}}{\partial \mathbf{\Sigma}}\frac{\partial}{\partial \mathbf{\Sigma}^{-1}}\text{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right) \\
&= -\frac{n}{2}\mathbf{\Sigma}^{-1} + \frac{n}{2}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{-1})\mathbf{S}^T \\
&= \frac{n}{2}\mathbf{\Sigma}^{-1}\left(-\mathbf{I} + \mathbf{\Sigma}^{-1}\mathbf{S}\right)
\end{aligned}$$

since in general

$$\frac{\partial}{\partial \mathbf{A}}\log|\mathbf{A}| = \mathbf{A}^{-1} \qquad \frac{\partial}{\partial \mathbf{A}}\text{tr}\left(\mathbf{B}\mathbf{A}\right) = \mathbf{B}^T \qquad \frac{\partial \mathbf{A}^{-1}}{\partial \mathbf{A}} = \mathbf{A}^{-1}\mathbf{A}^{-1}$$

This results into $\frac{\partial l}{\partial \mathbf{\Sigma}} = \mathbf{0}$ iff $\mathbf{\Sigma} = \mathbf{S}$

In the case of discrete (for example, binary) features we may simplify the model by assuming features are conditionally independent, given the class (naive Bayes hypothesis). Then,

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

where $p_{ki} = p(x_i = 1|C_k)$.

Functions $a_k(\mathbf{x})$ can then be defined as in the softmax model:

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

These are still linear functions on $\mathbf{x}$.

The property that $p(C_k|\mathbf{x})$ is a generalized linear model with sigmoid (for the binary case) and softmax (for the multiclass case) activation function holds more in general than assuming a gaussian or bernoulli class conditional distribution $p(\mathbf{x}|C_k)$.

Indeed, let the class conditional probability wrt C_k belong to the exponential family, that is it has the form

$$p(\mathbf{x}|\boldsymbol{\theta}_k) = g(\boldsymbol{\theta}_k)f(\mathbf{x})e^{\boldsymbol{\phi}(\boldsymbol{\theta}_k)^T\mathbf{u}(\mathbf{x})}$$

with the additional constraint that $\mathbf{u}$ is the identity function, that is $\mathbf{u}(\mathbf{x}) = \mathbf{x}$.

In the case of binary classification, we check that $a(\mathbf{x})$ is a linear function

$$a(\mathbf{x}) = \log \frac{p(\mathbf{x}|\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_1)}{p(\mathbf{x}|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2)} = \log \frac{g(\boldsymbol{\theta}_1)e^{\frac{1}{s}\boldsymbol{\phi}(\boldsymbol{\theta}_1)^T\mathbf{x}}p(\boldsymbol{\theta}_1)}{g(\boldsymbol{\theta}_2)e^{\frac{1}{s}\boldsymbol{\phi}(\boldsymbol{\theta}_2)^T\mathbf{x}}p(\boldsymbol{\theta}_2)}$$
$$= (\boldsymbol{\phi}(\boldsymbol{\theta}_1) - \boldsymbol{\phi}(\boldsymbol{\theta}_2))^T\,\mathbf{x} + \log g(\boldsymbol{\theta}_1) - \log g(\boldsymbol{\theta}_2) + \log p(\boldsymbol{\theta}_1) - \log p(\boldsymbol{\theta}_2)$$

Similarly, for multiclass classification, we may easily derive that

$$a_k(\mathbf{x}) = \boldsymbol{\phi}(\boldsymbol{\theta}_k)^T\mathbf{x} + \log g(\boldsymbol{\theta}_k) + p(\boldsymbol{\theta}_k)$$

for all k.

Probabilistic discriminative models

Generative models

For a large set of distributions type for $p(\mathbf{x}|C_k)$ the posterior class distributions $p(C_k|\mathbf{x})$ are sigmoidal (in the binary case) or softmax (for more classes): in both cases, with argument given by a linear combination of features in $\mathbf{x}$.

We may derive both the parameters of $p(\mathbf{x}|C_k)$ and the prior class probabilities $p(C_k)$ through maximum likelihood estimation, and next apply Bayes' rule to derive $p(C_k|\mathbf{x})$, at least up to a normalization factor.

Alternative idea

We could directly derive $p(C_k|\mathbf{x})$ (for example through ML estimation of its parameters).

Comparison wrt the generative approach:

- Less information derived (we do not know $p(\mathbf{x}|C_k)$, thus we are not able to generate new data)
- Simpler method, usually a smaller set of parameters to be derived
- Better predictions, if the assumptions done with respect to $p(\mathbf{x}|C_k)$ are poor.

A generalized linear model (GLM) is a function

$$y(\mathbf{x}) = f(\mathbf{w}^T\mathbf{x} + w_0)$$

where f is in general a non linear function.

Each iso-surface of $y(\mathbf{x})$, such that by definition $y(\mathbf{x}) = c$ (for some constant c), is such that

$$f(\mathbf{w}^T\mathbf{x} + w_0) = c$$

and

$$\mathbf{w}^T\mathbf{x} + w_0 = f^{-1}(y) = c'$$

(c' constant).

Hence, iso-surfaces of a GLM are hyper-planes, thus implying that boundaries are hyperplanes themselves.

Let us assume we wish to predict a random variable y as a function of a different set of random variables $\mathbf{x}$. A prediction model for this task is a GLM if the following hypotheses hold:

1. the conditional distribution of y given $\mathbf{x}$, $p(y|\mathbf{x})$ belongs to the exponential family (let $\boldsymbol{\theta}(\mathbf{x})$ be the corresponding natural parameter): that is,
$$p(y|\mathbf{x}) = g(\mathbf{x})f(y)e^{\boldsymbol{\theta}(\mathbf{x})^T\mathbf{u}(y)}$$
2. $E[y|\mathbf{x}]$ is considered as the prediction of y given $\mathbf{x}$
3. $\boldsymbol{\theta}(\mathbf{x})$ is a linear combination of the features,
$$\boldsymbol{\theta}(\mathbf{x}) = \mathbf{w}^T\overline{\mathbf{x}} = \sum_{i=1}^{D} w_i x_i + w_0$$

1. $y \in \mathbb{R}$, and $p(y|\mathbf{x}) \sim \mathcal{N}(y|\mu(\mathbf{x}), \sigma^2)$ is a normal distribution with mean $\mu(\mathbf{x})$ and constant variance σ^2: the natural parameter $\boldsymbol{\theta}(\mathbf{x})$ is, by definition,
$$\boldsymbol{\theta}(\mathbf{x}) = \left(\begin{array}{c} \theta_1(\mathbf{x}) \\ \theta_2 \end{array} \right) = \left(\begin{array}{c} \mu(\mathbf{x})/\sigma^2 \\ -1/2\sigma^2 \end{array} \right)$$
2. we wish to predict the value of y as $y(\mathbf{x}) = E[y|\mathbf{x}]$, then
$$y(\mathbf{x}) = \mu(\mathbf{x}) = \sigma^2\theta_1(\mathbf{x})$$
3. we assume there exists $\mathbf{w}$ such that $\theta_1(\mathbf{x}) = \mathbf{w}_1^T\overline{\mathbf{x}}$

Then, it results a linear regression
$$y(\mathbf{x}) = \mathbf{w}_1^T\overline{\mathbf{x}}$$

1. $y \in \{0, 1\}$, and $p(y|\mathbf{x}) \sim \mathcal{B}(y|\pi(\mathbf{x}))$ is a Bernoulli distribution with parameter $\pi(\mathbf{x})$: then, the natural parameter $\theta(\mathbf{x})$ is, by definition,

$$\theta(\mathbf{x}) = \log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}$$

2. we wish to predict the probability $p(y = 0|\mathbf{x})$ as

$$p(y = 0|\mathbf{x}) = E[y|\mathbf{x}] = \pi(\mathbf{x}) = \frac{1}{1 + e^{-\theta(\mathbf{x})}}$$

3. we assume there exists $\mathbf{w}$ such that $\theta = \mathbf{w}^T\overline{\mathbf{x}}$

Then, it results a logistic regression

$$p(y = 0|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\overline{\mathbf{x}}}}$$

- $y \in \{1, \ldots, K\}$, and $p(y|\mathbf{x}) \sim \mathcal{C}(y|\boldsymbol{\phi}(\mathbf{x}))$ is a categorical distribution with probabilities $\boldsymbol{\pi}(\mathbf{x}) = (\pi_1(\mathbf{x}), \ldots, \pi_K(\mathbf{x}))$ such that $\sum_{i=1}^K \pi_i(\mathbf{x}) = 1$ for all $\mathbf{x}$: the natural parameter $\boldsymbol{\theta}(\mathbf{x})$ of the distribution is, by definition, such that

$$\theta_i(\mathbf{x}) = \log \frac{\pi_i(\mathbf{x})}{\pi_K(\mathbf{x})} = \log \frac{\pi_i(\mathbf{x})}{1 - \sum_{j=1}^{K-1} \pi_j(\mathbf{x})}$$

- we wish to predict the probabilities $p(y = i|\mathbf{x})$ as

$$p(y = i|\mathbf{x}) = E[u_i(y)|\mathbf{x}] = \pi_i$$

where $\mathbf{u}(y)$ is the 1-to-K representation of y.
Then, it results $\pi_i(\mathbf{x}) = \pi_K(\mathbf{x})e^{\theta_i(\mathbf{x})}$ and, since $\sum_{i=1}^K \pi_i(\mathbf{x}) = 1$,

$$\pi_K(\mathbf{x}) = \frac{1}{\sum_{i=1}^K e^{\theta_i(\mathbf{x})}} \qquad \text{and} \qquad \pi_i(\mathbf{x}) = \frac{e^{\theta_i(\mathbf{x})}}{\sum_{i=1}^K e^{\theta_i(\mathbf{x})}}$$

- we assume there exists $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ such that $\theta_i(\mathbf{x}) = \mathbf{w}_i^T \overline{\mathbf{x}}$

Then, a softmax regression results, with

$$
\begin{aligned}
p(y = i|\mathbf{x}) &= \frac{e^{\mathbf{w}_i^T \overline{\mathbf{x}}}}{\sum_{j=1}^K e^{\mathbf{w}_j^T \overline{\mathbf{x}}}} \qquad \text{if } i \neq K \\
p(y = K|\mathbf{x}) &= \frac{1}{\sum_{j=1}^K e^{\mathbf{w}_j^T \overline{\mathbf{x}}}}
\end{aligned}
$$

As seen before, logistic regression is the GLM deriving from the hypothesis of a Bernoulli distribution of y, which results into

$$p(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})) = \frac{1}{1+e^{-\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})}}$$

where the use of basis functions is explicitly considered.

As observed, the model is equivalent, for the binary classification case, to linear regression for the regression case.

Degrees of freedom

- In the case of d features, logistic regression requires $d+1$ coefficients $w_0, \ldots, w_d$ to be derived from a training set
- A generative approach with gaussian distributions requires:
    - $2d$ coefficients for the means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$,
    - for each covariance matrix

$$\sum_{i=1}^{d} i = d(d+1)/2 \qquad \text{coefficients}$$

    - one prior cla probability $p(C_1)$
- As a total, it results into $d(d+1) + 2d + 1 = d(d+3) + 1$ coefficients (if a unique covariance matrix is assumed $d(d+1)/2 + 2d + 1 = d(d+5)/2 + 1$ coefficients)

Maximum likelihood estimation

- Training set $\mathbf{X}, \mathbf{t}$. The likelihood is

$$L(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \prod_{i=1}^{n} p_i^{t_i} (1 - p_i)^{1-t_i}$$

  where $p_i = p(C_1|\boldsymbol{\phi}(\mathbf{x}_i)) = \sigma(a_i)$, with $a_i = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i)$
- The log-likelihood is then

$$l(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \log L(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \sum_{i=1}^{n} (t_i \log p_i + (1 - t_i) \log(1 - p_i))$$

- Note that

$$\frac{\partial l(\mathbf{w}|\mathbf{X},\mathbf{t})}{\partial \mathbf{w}} = \sum_{i=1}^{n} \frac{\partial l(\mathbf{w}|\mathbf{X},\mathbf{t})}{\partial p_i} \frac{\partial p_i}{\partial a_i} \frac{\partial a_i}{\partial \mathbf{w}}$$

and

$$\frac{\partial l(\mathbf{w}|\mathbf{X},\mathbf{t})}{\partial p_i} = \frac{t_i}{p_i} - \frac{1-t_i}{1-p_i} = \frac{t_i - p_i}{p_i(1-p_i)}$$

$$\frac{\partial p_i}{\partial a_i} = \frac{\partial \sigma(a_i)}{\partial a_i} = \sigma(a_i)(1-\sigma(a_i)) = p_i(1-p_i)$$

$$\frac{\partial a_i}{\partial \mathbf{w}} = \boldsymbol{\phi}(\mathbf{x}_i)$$

- Hence,

$$\frac{\partial l(\mathbf{w}|\mathbf{X},\mathbf{t})}{\partial \mathbf{w}} = \sum_{i=1}^{n}(t_i - p_i)\boldsymbol{\phi}(\mathbf{x}_i) = \sum_{i=1}^{n}(t_i - \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)))\boldsymbol{\phi}(\mathbf{x}_i)$$

Maximum likelihood estimation

- To maximize the likelihood, we could apply a gradient ascent algorithm, where at each iteration the following update of the currently estimated $\mathbf{w}$ is performed

$$
\begin{aligned}
\mathbf{w}^{(j+1)} &= \mathbf{w}^{(j)} + \alpha \frac{\partial l(\mathbf{w}|\mathbf{X}, \mathbf{t})}{\partial \mathbf{w}}\Big|_{\mathbf{w}^{(j)}} \\
&= \mathbf{w}^{(j)} + \alpha \sum_{i=1}^{n} (t_i - \sigma((\mathbf{w}^{(j)})^T \boldsymbol{\phi}(\mathbf{x}_i)))\boldsymbol{\phi}(\mathbf{x}_i) \\
&= \mathbf{w}^{(j)} + \alpha \sum_{i=1}^{n} (t_i - y(\mathbf{x}_i))\boldsymbol{\phi}(\mathbf{x}_i)
\end{aligned}
$$

As a possible alternative, at each iteration only one coefficient in $\mathbf{w}$ is updated

$$
\begin{aligned}
w_k^{(j+1)} &= w_k^{(j)} + \alpha \frac{\partial l(\mathbf{w}|\mathbf{X}, \mathbf{t})}{\partial w_k}\Big|_{\mathbf{w}^{(j)}} \\
&= w_k^{(j+1)} + \alpha \sum_{i=1}^{n} (t_i - \sigma((\mathbf{w}^{(j)})^T \boldsymbol{\phi}(\mathbf{x}_i)))\phi_k(\mathbf{x}_i) \\
&= w_k^{(j+1)} + \alpha \sum_{i=1}^{n} (t_i - y(\mathbf{x}_i))\phi_k(\mathbf{x}_i)
\end{aligned}
$$

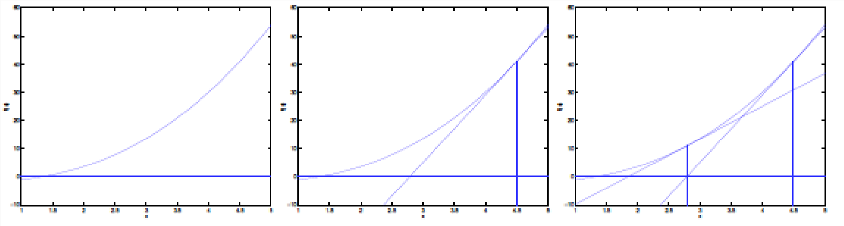
- Maximization of $l(\mathbf{w}|\mathbf{X}, \mathbf{t})$ through the well-known Newton-Raphson algorithm to compute the roots of a given function
- Given $f : \mathbb{R} \mapsto \mathbb{R}$, the algorithm finds $z \in \mathbb{R}$ such that $f(z) = 0$ through a sequence of iterations, starting from an initial value z_0 and performing the following update

$$z_{i+1} = z_i - \frac{f(z_i)}{f'(z_i)}$$

- At each iteration, the algorithm approximates f by a tangent line to f in $(z_i, f(z_i))$ and tangent to f, and defines z_{i+1} as the value where the line intersects the x axis

Newton-Raphson method

- Example of application of the method



- Newton-Raphson method can be also applied to compute maximum and minimum points for a function by finding zeros of the first derivative: this corresponds to applying the following update

$$z_{i+1} = z_i - \frac{f'(z_i)}{f''(z_i)}$$

Newton-Raphson and multivariate functions

- To apply Newton-Raphson to logistic regression we have to extend it to the case of a vector variable, since the maximization has to be performed with respect to the vector $\mathbf{w}$ of coefficients
- In a multivariate framework, the first derivative is substituted by the gradient $\frac{\partial}{\partial \mathbf{w}}$, while the second derivative corresponds to the Hessian matrix $\mathbf{H}$, defined as follows

$$\mathbf{H}_{ij}(f) = \frac{\partial^2 f}{\partial w_i \partial w_j}$$

- The update operation turns out to be

$$\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} - (\mathbf{H}(f)|_{\mathbf{w}^{(i)}})^{-1} \frac{\partial f}{\partial \mathbf{w}}\Big|_{\mathbf{w}_{(i)}}$$

Newton-Raphson and linear regression

- The error function, to be minimized, is

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) - t_i)^2$$

- Then,

$$\frac{\partial E}{\partial \mathbf{w}} = \sum_{i=1}^{n}(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) - t_i)\boldsymbol{\phi}(\mathbf{x}_i) = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Phi}^T\mathbf{t}$$

$$\mathbf{H} = \frac{\partial}{\partial \mathbf{w}}\frac{\partial E}{\partial \mathbf{w}} = \sum_{i=1}^{n}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

- At each iteration, the update is

$$\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} - (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w}^{(i)} - \boldsymbol{\Phi}^T\mathbf{t}) = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{t}$$

- We obtain the well-known solution, which is obtained in a single iteration.

Here, we have

$$E(\mathbf{w}) = -l(\mathbf{w}|\mathbf{X},\mathbf{t}) = -\sum_{i=1}^{n}\Big(t_i \ln\sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) + (1-t_i)\ln(1-\sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)))\Big)$$

(this is called cross-entropy function). Hence,

$$\frac{\partial E}{\partial \mathbf{w}} = -\frac{\partial l(\mathbf{w}|\mathbf{X},\mathbf{t})}{\partial \mathbf{w}} = \sum_{i=1}^{n}(\sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) - t_i)\boldsymbol{\phi}(\mathbf{x}_i) = \boldsymbol{\Phi}^T(\mathbf{s}_\mathbf{w} - \mathbf{t})$$

$$\mathbf{H} = \frac{\partial}{\partial \mathbf{w}}\frac{\partial E}{\partial \mathbf{w}} = \sum_{i=1}^{n}\sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))(1-\sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)))\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T = \boldsymbol{\Phi}^T\mathbf{R}_\mathbf{w}\boldsymbol{\Phi}$$

where

- $\mathbf{s}_\mathbf{w}$ is a vector such that $\mathbf{s}_{\mathbf{w}i} = \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))$ for $i = 1, \ldots, n$
- $\mathbf{R}_\mathbf{w}$ is a diagonal matrix such that

$$\mathbf{R}_{\mathbf{w}ii} = \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))(1-\sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))) = \mathbf{s}_{\mathbf{w}i}(1-\mathbf{s}_{\mathbf{w}i})$$

- In the case of logistic regression, the update is then

$$\begin{aligned}\mathbf{w}^{(i+1)} &= \mathbf{w}^{(i)} - (\mathbf{\Phi}^T\mathbf{R}_{\mathbf{w}^{(i)}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T(\mathbf{s}_{\mathbf{w}^{(i)}} - \mathbf{t}) \\ &= (\mathbf{\Phi}^T\mathbf{R}_{\mathbf{w}^{(i)}}\mathbf{\Phi})^{-1}((\mathbf{\Phi}^T\mathbf{R}_{\mathbf{w}^{(i)}}\mathbf{\Phi})\mathbf{w}^{(i)} - \mathbf{\Phi}^T(\mathbf{s}_{\mathbf{w}^{(i)}} - \mathbf{t})) \\ &= (\mathbf{\Phi}^T\mathbf{R}_{\mathbf{w}^{(i)}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{R}_{\mathbf{w}^{(i)}}\mathbf{z}_{\mathbf{w}^{(i)}}\end{aligned}$$

where $\mathbf{z}_{\mathbf{w}^{(i)}}$ is a vector of size d defined as

$$\mathbf{z}_{\mathbf{w}^{(i)}} = \mathbf{\Phi}\mathbf{w}^{(i)} - \mathbf{R}^{-1}_{\mathbf{w}^{(i)}}(\mathbf{s}_{\mathbf{w}^{(i)}} - \mathbf{t})$$

As can be seen, $\mathbf{z}_{\mathbf{w}^{(i)}}$ is a function of $\mathbf{w}^{(i)}$, hence of i.

Iterated reweighted least squares

- The value $(\mathbf{\Phi}^T\mathbf{R}_{\mathbf{w}^{(i)}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{R}_{\mathbf{w}^{(i)}}\mathbf{z}_{\mathbf{w}^{(i)}}$ can be seen as the solution of a suitable instance of the weighted least squares problem defined as the minimization of
$$\sum_{i=1}^{n}\psi_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)-t_i)^2$$
for given weights $\psi_1,\ldots,\psi_n$
- The minimum of this problem is obtained for
$$\mathbf{w}=(\mathbf{\Phi}^T\mathbf{\Psi}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{\Psi}\mathbf{t}$$
where $\mathbf{\Psi}$ is a diagonal matrix such that $\mathbf{\Psi}_{ii}=\psi_i$
- In our case $\mathbf{\Psi}=\mathbf{R}_{\mathbf{w}^{(i)}}$ and $\mathbf{t}=\mathbf{z}_{\mathbf{w}^{(i)}}=\mathbf{\Phi}\mathbf{w}^{(i)}-\mathbf{R}^{-1}_{\mathbf{w}^{(i)}}(\mathbf{s}_{\mathbf{w}^{(i)}}-\mathbf{t})$: both of them are functions of i
- The update of $\mathbf{w}^{(i)}$ performed at each iteration implies solving a new instance of the weighted least square problem, setting $\mathbf{w}^{(i+1)}$ to the solution obtained, and deriving the new values $\mathbf{R}_{\mathbf{w}^{(i+1)}}$ and $\mathbf{z}_{\mathbf{w}^{(i+1)}}$.

- Observe that assuming $p(\mathbf{x}|C_1)$ are $p(\mathbf{x}|C_2)$ as multivariate normal distributions with same covariance matrix $\mathbf{\Sigma}$ results into a logistic $p(C_1|\mathbf{x})$.
- The opposite, however, is not true in general: in fact, GDA relies on stronger assumptions than logistic regression.
- The more the normality hypothesis of class conditional distributions with same covariance is verified, the more GDA will tend to provide the best models for $p(C_1|\mathbf{x})$

- Logistic regression relies on weaker assumptions than GDA: it is then less sensible from a limited correctness of such assumptions, thus resulting in a more robust technique
- Since $p(C_i|\mathbf{x})$ is logistic under a wide set of hypotheses about $p(\mathbf{x}|C_i)$, it will usually provide better solutions (models) in all such cases, while GDA will provide poorer models as far as the normality hypotheses is less verified.

- In order to extend the logistic regression approach to the case $K > 2$, let us consider the vector $\mathbf{w}$ of model coefficients, of size dK, where the k-th block of $\mathbf{w}$ $(k = 1, \dots, K)$ corresponds to the vector $\mathbf{w}_k$ of coefficients for class C_k.
- In this case, the likelihood is defined as

$$
\begin{aligned}
p(\mathbf{T}, \mathbf{X}|\mathbf{w}) &= \prod_{i=1}^{n} \prod_{k=1}^{K} p(C_k|\mathbf{x}_i)^{t_{ik}} \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{e^{\mathbf{w}_k^T \boldsymbol{\phi}(\mathbf{x}_i)}}{\sum_{r=1}^{K} e^{\mathbf{w}_r^T \boldsymbol{\phi}(\mathbf{x}_i)}} \right)^{t_{ik}}
\end{aligned}
$$

  where $\mathbf{X}$ is the usual matrix of features and $\mathbf{T}$ is an $n \times K$ matrix such that the i-th row of $\mathbf{T}$ is the 1-to-K coding of t_i. That is, if $\mathbf{x}_i \in C_k$ then $t_{ik} = 1$ and $t_{ir} = 0$ for $r \neq k$.

The log-likelihood is then defined as

$$l(\mathbf{w}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \log \left(\frac{e^{\mathbf{w}_k^T \boldsymbol{\phi}(\mathbf{x}_i)}}{\sum_{r=1}^{K} e^{\mathbf{w}_r^T \boldsymbol{\phi}(\mathbf{x}_i)}} \right)$$

The gradient $\dfrac{\partial l(\mathbf{w})}{\partial \mathbf{w}}$ is a vector of size dK, where its j-th block $(j = 1, \ldots, K)$ corresponds to $\dfrac{\partial l(\mathbf{w})}{\partial \mathbf{w}_j}$.

- To derive $\dfrac{\partial l(\mathbf{w})}{\partial \mathbf{w}_j}$ let

$$y_{ik} = \frac{e^{a_{ik}}}{\sum_{r=1}^{K} e^{a_{ir}}} \qquad \text{with} \qquad a_{ik} = \mathbf{w}_k^T \boldsymbol{\phi}(\mathbf{x}_i)$$

  for $k = 1, \dots, K$ and $i = 1, \dots, n$. Then,

$$l(\mathbf{w}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \log y_{ik}$$

- For each $i = 1, \dots, n$, $j = 1, \dots, M$, $k = 1, \dots, K$,

$$\begin{aligned}
\frac{\partial a_{ik}}{\partial w_{kj}} &= \frac{\partial}{\partial w_{kj}} \mathbf{w}_k^T \boldsymbol{\phi}(\mathbf{x}_i) = \phi_j(\mathbf{x}_i) \\
\frac{\partial y_{ik}}{\partial a_{ik}} &= y_{ik}(1 - y_{ik}) \\
\frac{\partial y_{ik}}{\partial a_{ir}} &= -y_{ir} y_{ik} \qquad \text{if } r \neq k
\end{aligned}$$

Hence,

$$
\begin{aligned}
\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}_j} &= \frac{\partial}{\partial \mathbf{w}_j}\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}\log y_{ik} = \frac{\partial l}{\partial \mathbf{w}_j}\sum_{i=1}^{n} t_{ij}\log y_{ij} + \frac{\partial l}{\partial \mathbf{w}_j}\sum_{k\neq j}\sum_{i=1}^{n} t_{ik}\log y_{ik} \\
&= \sum_{i=1}^{n} t_{ij}\frac{1}{y_{ij}}\frac{\partial y_{ij}}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial \mathbf{w}_j} + \sum_{k\neq j}\sum_{i=1}^{n} t_{ik}\frac{1}{y_{ik}}\frac{\partial y_{ik}}{\partial a_{ik}}\frac{\partial a_{ik}}{\partial \mathbf{w}_j} \\
&= \sum_{i=1}^{n} t_{ij}\frac{1}{y_{ij}} y_{ij}(1-y_{ij})\boldsymbol{\phi}(\mathbf{x}_i) - \sum_{k\neq j}\sum_{i=1}^{n} t_{ik}\frac{1}{y_{ik}} y_{ik}y_{ij}\boldsymbol{\phi}(\mathbf{x}_i) \\
&= \left(\sum_{i=1}^{n} t_{ij} - \sum_{i=1}^{n} y_{ij}\sum_{k=1}^{K} t_{ik}\right)\boldsymbol{\phi}(\mathbf{x}_i) \\
&= \left(\sum_{i=1}^{n} t_{ij} - \sum_{i=1}^{n} y_{ij}\right)\boldsymbol{\phi}(\mathbf{x}_i) = \sum_{i=1}^{n}(t_{ij}-y_{ij})\boldsymbol{\phi}(\mathbf{x}_i)
\end{aligned}
$$

Observe that the gradient has the same structure than in the case of linear regression and logistic regression.

Probit regression

- In a GLM, $p(\mathcal{C}_1|\mathbf{x}) = f(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))$ where f is the activation function (a sigmoid in the case of logistic regression)
- In probit regression a stochastic threshold model is applied for classification, as follows:
    - Let $\mathbf{w}$ be the model coefficients. In order to classify $\mathbf{x}_i$, the linear combination $a_i = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)$ is computed
    - A threshold value θ is sampled from a given distribution $p(\theta)$
    - $\mathbf{x}_i$ is classified in $\mathcal{C}_1$ if $a_i \geq \theta$, otherwise it is classified in $\mathcal{C}_0$.
- In this case, we identify the activation function as the probability that $\mathbf{x}_i$ is classified in $\mathcal{C}_1$, which is given by the cumulative function

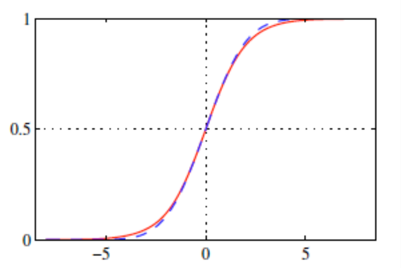
$$f(a) = \int_{-\infty}^{a} p(\theta)d\theta$$

Probit regression

- A relevant case is the one of a gaussian $p(\theta)$ with zero mean and unitary variance, which results into a probit activation function

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} d\theta$$

- observe that $\Phi(a)$ is monotonically increasing, with $0 < \Phi(a) < 1$



- Assuming a more general gaussian distribution for $p(\theta)$ does not change the model, since it is possible to prove that this corresponds to a rescaling of the coefficients in $\mathbf{w}$.

Bayesian logistic regression

Bayesian logistic regression

- Used to overcome the overfitting problem by assuming a prior distribution
- The aim is to estimate the posterior class distribution

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}, \mathbf{X}, \mathbf{t}) &= \int p(\mathcal{C}_1|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{t}) d\mathbf{w} \\ &= \int \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})) p(\mathbf{w}|\mathbf{X}, \mathbf{t}) d\mathbf{w} \end{aligned}$$

- Thus, we need to derive the posterior distribution of coefficients $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$: this is in general intractable

By Bayes' rule,

$$p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})} = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{t}|\mathbf{X},\mathbf{w}')p(\mathbf{w}')d\mathbf{w}'}$$

where the likelihood is $p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{i=1}^{n} p(t_i|\mathbf{x}_i,\mathbf{w})$, with

$$p(t_i|\mathbf{x}_i,\mathbf{w}) = \begin{cases} \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})) & \text{if } t_i = 1 \\ 1 - \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})) & \text{if } t_i = 0 \end{cases}$$

that is,

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{i=1}^{n} \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))^{t_i} \left(1 - \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))\right)^{1-t_i}$$

and

$$p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{w}) \prod_{i=1}^{n} \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))^{t_i} \left(1 - \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))\right)^{1-t_i}}{Z}$$

with

$$Z = \int p(\mathbf{w}) \prod_{i=1}^{n} \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))^{t_i} \left(1 - \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))\right)^{1-t_i} d\mathbf{w}$$

Predictive distribution intractability

Since the predictive distribution is the expectation of the model prediction wrt to the distribution of model coefficients,

$$p(\mathcal{C}_1|\mathbf{x}, \mathbf{X}, \mathbf{t}) = \int \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))p(\mathbf{w}|\mathbf{X}, \mathbf{t})d\mathbf{w}$$

we need some way to evaluate $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$ for any $\mathbf{w}$. Unfortunately, since Z is hard to compute, we are only able to evaluate

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}) = p(\mathbf{w}) \prod_{i=1}^{n} \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))^{t_i} \left(1 - \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))\right)^{1-t_i}$$

which is proportional to $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$ through an unknown proportionality coefficient.

Possible options:

1. find a single value of $\mathbf{w}$ which maximizes $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$: this corresponds to the value which maximizes $g(\mathbf{w}; \mathbf{X}, \mathbf{t})$ (this is the usual MAP approach)
2. approximate $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$ with some other probability density which can be treated analytically
3. sample from $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$, knowing only $g(\mathbf{w}; \mathbf{X}, \mathbf{t})$

Bayesian logistic regression: MAP

In order to approximately estimate the posterior distribution $p(\mathbf{w}|\mathbf{X},\mathbf{t}) \propto p(\mathbf{X},\mathbf{t}|\mathbf{w})p(\mathbf{w})$ we assume a simple gaussian prior with mean $\mathbf{0}$ and diagonal covariance $\sigma^2\mathbf{I}$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\sigma^2) = \frac{1}{(2\pi)^{\frac{D}{2}}\sigma^D} e^{-\frac{\mathbf{w}^T\mathbf{w}}{2\sigma^2}}$$

Since the training set likelihood wrt the parameter is, as usual,

$$p(\mathbf{X},\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{n} y_i^{t_i}(1-y_i)^{1-t_i}$$

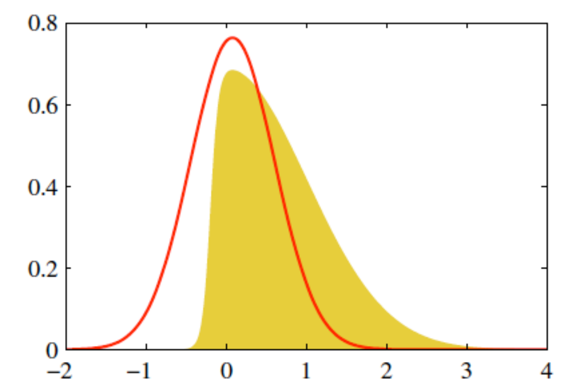
where $y_i = y(\mathbf{x}_i,\mathbf{w}) = \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))$, the logarithm of the posterior results as follows

$$\log p(\mathbf{w}|\mathbf{X},\mathbf{t}) = -\frac{1}{2\sigma^2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n}(t_i \log y_i + (1-t_i)\log(1-y_i)) + c$$

The MAP value for $\mathbf{w}$ can be found, for example, by applying Newton-Raphson (the gradient and the Hessian matrix can be easily derived)

Bayesian logistic regression: Laplace approximation

- A distribution $p(\mathbf{z})$ is approximated by a gaussian $q(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean corresponding to a maximum of $p(\mathbf{z})$ and variance equal to some suitable value



- Let $p(z) = \frac{1}{Z} f(z)$, where $Z = \int f(z)dz$ is the (unknown) normalization coefficient
- A mode (maximum probability value) z_0 of $p(z)$ (and $f(z)$) can be found by solving
$$0 = \frac{df(z)}{dz}$$
- A gaussian distribution has one single mode, corresponding to its mean μ. In Laplace approximation, we set the mean of $q(z)$ equal to the mode of $f(z)$, that is we assume $\mu = z_0$
- Moreover, $\log q(z)$ is a quadratic function
$$\log q(z) = -\frac{1}{2\sigma^2}(z - \mu)^2 + c$$
It will be approximated by means of another suitable quadratic functions

Laplace approximation for $d = 1$

- Consider the first two terms of the Taylor series expansion of $\log f(z)$ around z_0

$$\log f(z) \simeq \log f(z_0) + \left.\frac{d \log f(z)}{dz}\right|_{z=z_0} (z - z_0) + \frac{1}{2} \left.\frac{d^2 \log f(z)}{dz^2}\right|_{z=z_0} (z - z_0)^2$$

- Since z_0 corresponds to a maximum of $f(z)$, and also of $\log f(z)$, we have $\left.\frac{d \log f(z)}{dz}\right|_{z=z_0} = 0$, hence

$$\log f(z) \simeq \log f(z_0) - \frac{1}{2} A (z - z_0)^2$$

where

$$A = -\left.\frac{d^2 \log f(z)}{dz^2}\right|_{z=z_0}$$

- By comparing approximations for $\log q(z)$ and $\log f(z)$ we get

$$\sigma^2 = \frac{1}{A}$$

Overall, we obtain

$$q(z) = Ce^{-\frac{A}{2}(x-z_0)^2}$$

C is the normalization factore of a gaussian. Then,

$$q(z) = \frac{\sqrt{A}}{\sqrt{2\pi}}e^{-\frac{A}{2}(x-z_0)^2}$$

Laplace approximation for $d > 1$

We make the same considerations as in the 1-dimensional case:

- Let $\mathbf{z}_0$ be such that $\left.\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}}\right|_{\mathbf{z}=\mathbf{z}_0} = 0$
- Consider the Taylor expansion around $\mathbf{z}_0$

$$\log f(\mathbf{z}) \simeq \log f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A}(\mathbf{z}-\mathbf{z}_0)$$

  where $\mathbf{A}$ is the Hessian matrix of $-\log f(\mathbf{z})$, computed in $\mathbf{z}_0$:

$$\mathbf{A} = -\left.\frac{\partial^2 \log f(\mathbf{z})}{\partial \mathbf{z}^2}\right|_{\mathbf{z}=\mathbf{z}_0} = \mathbf{H}(-\log f(\mathbf{z}))\Big|_{\mathbf{z}=\mathbf{z}_0}$$

- Then, $f(\mathbf{z}) \simeq f(\mathbf{z}_0)e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A}(\mathbf{z}-\mathbf{z}_0)}$ around $\mathbf{z}_0$. By setting the covariance matrix $\mathbf{\Sigma}$ of $q(\ z)$ equal to $\mathbf{A}^{-1}$, we get

$$q(\mathbf{z}) = Ce^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A}(\mathbf{z}-\mathbf{z}_0)}$$

  which, by normalizing, results into

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A}(\mathbf{z}-\mathbf{z}_0)}$$

If we apply Laplace approximation to $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$, we get a gaussian distribution $p(\mathbf{w}|\overline{\mathbf{w}}, \overline{\mathbf{\Sigma}})$ such that

- $\overline{\mathbf{w}} = \mathbf{w}_{MAP}$ is computed as sketched before
- $\overline{\mathbf{\Sigma}}$ is defined as

$$\overline{\mathbf{\Sigma}} = -\frac{\partial^2}{\partial \mathbf{w}^2} \log p(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \mathbf{\Sigma}_0^{-1} + \sum_{i=1}^{n} y_i(1 - y_i)\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T$$

However, the expectation for the predictive distribution

$$p(\mathcal{C}_1|\mathbf{x}, \mathbf{X}, \mathbf{t}) = \int \sigma(\mathbf{w}^T \Phi(\mathbf{x}))p(\mathbf{w}|\mathbf{X}, \mathbf{t})d\mathbf{w} \simeq \int y(\mathbf{x}, \mathbf{w})\mathcal{N}(\mathbf{w}|\overline{\mathbf{w}}, \overline{\mathbf{\Sigma}})d\mathbf{w}$$

can still be impossible to deal with analytically

Possible approaches:

- apply some further approximation. Under suitable assumptions this leads to

$$p(\mathcal{C}_1|\mathbf{x}, \mathbf{X}, \mathbf{t}) \simeq \sigma\left(\frac{\mathbf{w}_{MAP}^T \boldsymbol{\phi}(\mathbf{x})}{\sqrt{1 + \frac{\pi}{8}\boldsymbol{\phi}(\mathbf{x})^T \overline{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\mathbf{x})}}\right)$$

- sample a set of N_s values $\mathbf{w}_i$ from $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$; evaluate the corresponding model $\sigma(\mathbf{w}_i^T \boldsymbol{\phi}(\mathbf{x}))$ for each sampled value $\mathbf{w}_i$; approximate the expectation by means of the average on the set of sampled values

$$p(\mathcal{C}_1|\mathbf{x}, \mathbf{X}, \mathbf{t}) \simeq \frac{1}{N_s} \sum_{i=1}^{N_s} \sigma(\mathbf{w}_i^T \boldsymbol{\phi}(\mathbf{x}))$$

In this case, we still apply the approximation

$$p(\mathcal{C}_1|\mathbf{x}, \mathbf{X}, \mathbf{t}) \simeq \frac{1}{N_s}\sum_{i=1}^{N_s}\sigma(\mathbf{w}_i^T\boldsymbol{\phi}(\mathbf{x}))$$

where the set of value is now sampled from $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$ and not from its approximation.

Some Markov Chain Montecarlo (MCMC) sampling method can be applied, such as Metropolis-Hastings or Gibbs sampling