Logistic regression

Logistic regression is used for binary classification.
It is quite similar to a simple linear regression in the sense that the objective is to find optimal weights ω to predict a variable. However, in the logistic regression we use a sigmoïd function.

Rem: "logistic" because the logistic law has a sigmoïd function as a repartition function.

Rationale behind the use of the sigmoïd function:
We look for the *à posteriori* probability $\mathbb{P}(y=1|x)=\pi(x)=\hat{y}$.
The predicted variable $\hat{y}$ is thus a probability.

The sigmoïd function: $\sigma : z \rightarrow \frac{1}{1+e^{-z}}$ is well adapted because of two reasons:
1) We want an output variable that is included in $[0,1]$
2) $\frac{\pi(z)}{1-\pi(z)}$ represents the relationship between a distribution and its complementary (good in binary case), and it is just a transformation of $\sigma(z)=\frac{1}{1+e^{-z}}=\frac{e^z}{1-e^z}$

Thus, we have:
$\hat{y}=\mathbb{P}(y=1|x)=\sigma(\omega^T x+b)=\frac{1}{1-e^{-(\omega^T x+b)}}$

Estimation
Estimation is done using maximum likelihood. Maximum likelihood is finding the parameter that maximizes the probability to have a specific event (x_i,y_i) but in our case, it is a *conditional* maximum likelihood since we want to maximize the *à posteriori* probability that depends on x.

$L(\omega,b)=\prod_{i=1}^{n}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$

This equation has no analytic solution. We use a numeric method to find the optimal parameters (see optimizaton algorithms).

Expectation-Maximization (EM) in the case of GMM (Gaussian Mixture Model)

(for more details, see document *gmm.pdf* in Cloud folder)

A GMM sample is composed of j Gaussian variables (*clusters*) distributed with proportions $(\pi_1,...,\pi_k)$ $(\Sigma\pi_i=1)$
We can write:

$$X \sim \mathcal{N}(\mu_Z,\Sigma_Z) \qquad with \ \ Z \sim \pi$$

π is not really a law but more the proportions of each Gaussian categories.
Thus, X has a density which is a weighted-average of all Gaussian densities:

$$p_\theta(x)=\Sigma_{j=1}^{k}\pi_j f_j(x) \qquad (*)$$

Estimation
We want to estimate $\theta=(\pi,\mu,\Sigma)$ where:

$\pi = (\pi_1, ..., \pi_k)$, $\mu = (\mu_1, ..., \mu_k)$, $\Sigma = (\Sigma_1, ..., \Sigma_k)$

To do so, we use the maximum likelihood method (product of densities across all samples):

$$p_\theta(x) = \Pi_{i=1}^n p_\theta(x_i)$$

$$l(\theta) = log(\Pi_{i=1}^n p_\theta(x_i)) = \Sigma_{i=1}^n log(p_\theta(x_i))$$

We thus need to find $argmax(l(\theta))$
Problem: the likelihood function is not convex!

The expectation-maximization problem is used when we have *latent variables* (= variables for which we don't know their associated distribution).
Let $z = (z_1, ..., z_k)$ be the vector of latent variables. We can express the density $(*)$ as a joint function with respect to z:

$$p_\theta(x, z) = p_\theta(z)p_\theta(x|z)$$

$$l(\theta, z) = ... = \Sigma(log\pi_{z_i}) + \Sigma(logf_{z_i}(x_i))$$

A classic optimization (in case of Gaussians) give us empirical values as solutions e.g. $\hat{\pi_j} = \frac{n_j}{n}$
Problem: we don't know j!

We will thus use the *expected* log-likelihood method.
Let us find another expression of the likelihood:

$$p_\theta(x, z) = p_\theta(x)p_\theta(z|x)$$

As seen previously: $p_\theta(x, z) = \Pi\pi_{z_i}f_{z_i}(x_i)$
$p_\theta(z|x) = \Pi p_\theta(z_i|x_i) = \frac{\Pi\pi_{z_i}fz_i(x_i)}{p_\theta(x_i)} \propto \Pi\pi_{z_i}fz_i(x_i)$

Given an initial parameter θ_0, the *expected* log-likelihood is written as such:

$$\mathbb{E}_{\theta_0}[l(\theta; z)] = \Sigma p_{\theta_0}(z|x)l(\theta; z)$$
$$\mathbb{E}_{\theta_0}[l(\theta; z)] = \Sigma_j\Sigma_i p_{ij}(log\pi_j + logf_j(x_i))$$

We now have an expression that doesn't depend on z but only on p_{ij} and we know that $n_j = \Sigma_i p_{ij}$