MAP vs. Logistic Regression

Consider a binary classification problem described by two random variables, the hypothesis $Y \in \{-1, 1\}$ and the observation $X \in \mathbb{R}$.

Case 1: perfectly known statistical model.

The two hypotheses are equally probable a priori, that is:

$$\pi(-1) = \mathbb{P}[Y = -1] = \frac{1}{2}, \qquad \pi(+1) = \mathbb{P}[Y = +1] = \frac{1}{2}. \tag{1}$$

The conditional distribution of the feature X given Y is Gaussian, with variance equal to σ^2 and a mean that depends on the hypothesis, specifically:

$$\begin{aligned} \ell(x|Y = -1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x+1)^2}{2\sigma^2}\right\}, \\ \ell(x|Y = +1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-1)^2}{2\sigma^2}\right\}. \end{aligned} \tag{2}$$

1. Compute the posterior pmf, $p(y|x) = \mathbb{P}[Y = y|X = x]$.
2. Choose two values for σ^2, denoted as σ^2_{easy} and σ^2_{diff}, which should respectively represent an "easy" and a "difficult" classification problem. (**It is recommended to choose variance values that: i) are not "extreme", in order to avoid error probabilities that are too small or too close to 1/2; and ii) are sufficiently different in order to highlight the differences between the two scenarios**). Graphically represent using a computer the function $p(+1|x)$ as x varies, for the two selected variance values. Comment on the result, relating the shape of the plotted curves to the difficulty of the classification problem.
3. Empirically evaluate, via Monte Carlo simulation, the error probability of the MAP method for the two chosen variance values, and comment on the result.

Case 2: supervised classification.

Now generate a training set assuming the model described above, using only the value of variance σ^2_{easy}. The student is free to select a sufficiently large number of examples to ensure good performance of the learning algorithms to be implemented below.

1. Use the logistic regression method for binary classification, training the system with a stochastic gradient algorithm.
2. Using the parameters estimated in the previous step, empirically evaluate the performance (in terms of error probability) of the classifier obtained in the previous step. Compare the results with the known model case and provide adequate commentary.