MAP vs. Regressione Logistica

Si consideri un problema di classificazione binario descritto da due variabili aleatorie, l'ipotesi $Y \in \{-1, 1\}$ e l'osservazione $X \in \mathbb{R}$.

Caso 1: modello statistico perfettamente noto.

Le due ipotesi sono equiprobabili a priori, vale a dire:

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

La distribuzione condizionata delle feature X dato Y è Gaussiana, con varianza pari a σ^2 e media che dipende dall'ipotesi, specificamente:

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

1. Si calcoli la pmf a posteriori, $p(y|x) = \mathbb{P}[Y = y|X = x]$.
2. Si scelgano poi due valori per σ^2, indicati come σ^2_{easy} ed σ^2_{diff}, che siano rispettivamente rappresentativi di un problema di classificazione "facile" e di uno "difficile". (**Si consiglia di scegliere valori di varianza: i) non "estremi", in modo da evitare probabilità di errore troppo piccole o troppo prossime a 1/2; e ii) sufficientemente diversi in modo da evidenziare le differenze tra i due scenari**). Si rappresenti graficamente al calcolatore la funzione $p(+1|x)$ al variare di x, per i due valori di varianza scelti. Si commenti il risultato ottenuto, mettendo in relazione la forma delle curve rappresentate e la difficoltà del problema di classificazione.
3. Si valuti empiricamente, attraverso simulazione Monte Carlo, la probabilità di errore del metodo MAP per i due valori di varianza scelti, e si commenti il risultato.

Caso 2: classificazione supervisionata.

Si generi ora un training set assumendo il modello sopra descritto, per il solo valore di varianza σ^2_{easy}. Lo studente è libero di selezionare un numero di esempi sufficientemente grande da garantire buone prestazioni degli algoritmi di apprendimento da implementare nel seguito.

1. Utilizzare il metodo della regressione logistica per la classificazione binaria, addestrando il sistema con un algoritmo del gradiente stocastico.
2. Utilizzando i parametri stimati al punto precedente, calcolare empiricamente le prestazioni (in termini di probabilità di errore) del classificatore ottenuto al punto precedente. Confrontare i risultati ottenuti con il caso di modello noto e commentare adeguatamente.