

1.2.3 Modèles linéaires généralisés

Famille exponentielle – Une classe de distributions est issue de la famille exponentielle lorsqu'elle peut être écrite en termes d'un paramètre naturel, aussi appelé paramètre canonique ou fonction de lien η , d'une statistique suffisante $T(y)$ et d'une fonction de log-partition $a(\eta)$ de la manière suivante :

$$p(y; \eta) = b(y) \exp(\eta T(y) - a(\eta))$$

Remarque : on aura souvent $T(y) = y$. Aussi, $\exp(-a(\eta))$ peut être vu comme un paramètre de normalisation s'assurant que les probabilités somment à un.

Les distributions exponentielles les plus communément rencontrées sont récapitulées dans le tableau ci-dessous :

Distribution	η	$T(y)$	$a(\eta)$	$b(y)$
Bernoulli	$\log\left(\frac{\phi}{1-\phi}\right)$	y	$\log(1 + \exp(\eta))$	1
Gaussian	μ	y	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$
Poisson	$\log(\lambda)$	y	e^η	$\frac{1}{y!}$
Geometric	$\log(1 - \phi)$	y	$\log\left(\frac{e^\eta}{1 - e^\eta}\right)$	1

Hypothèses pour les GLMs – Les modèles linéaires généralisés (GLM) ont pour but de prédire une variable aléatoire y comme une fonction de $x \in \mathbb{R}^{n+1}$ et reposent sur les 3 hypothèses suivantes :

$$(1) \quad y|x; \theta \sim \text{ExpFamily}(\eta) \quad (2) \quad h_\theta(x) = E[y|x; \theta] \quad (3) \quad \eta = \theta^T x$$

Remarque : la méthode des moindres carrés ordinaires et la régression logistique sont des cas spéciaux des modèles linéaires généralisés.

1.3 Support Vector Machines

Le but des support vector machines est de trouver la ligne qui maximise la distance minimum à la ligne.

Classifieur à marges optimales – Le classifieur à marges optimales h est tel que :

$$h(x) = \text{sign}(w^T x - b)$$

où $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ est une solution du problème d'optimisation suivant :

$$\min \frac{1}{2} \|w\|^2 \quad \text{tel que} \quad y^{(i)} (w^T x^{(i)} - b) \geq 1$$