Inférence Bayesienne
Cours 1

1 Classification Baysienne

Y: la classe à predire (catégorique)
$\vec{X}$: vecteur aléatoire, $\vec{X} = vec(X_1, ...X_d)$

Choisir y qui maximise

$$P(Y = y|\vec{X} = \vec{x}) = \left.\frac{\overbrace{P(\vec{X} = \vec{x}|Y = y)}^{\text{vraisemblence}}\overbrace{P(Y = y)}^{\text{à priorie}}}{\underbrace{P(\vec{X} = \vec{x})}_{\text{évidence}}}\right] \text{(niveau 1)}$$

à estimé: $P(Y)$, $P(\vec{X}|Y) \rightarrow$ Pour chaque classe y une distribution sur

$$\vec{X} \xrightarrow[\text{(Hypothese naïve)}]{} P(\vec{X} = \vec{x}) = \prod_{i=d}^{d} \underbrace{P(X_i = x_i|Y = y)}$$

Bernoulli $\rightarrow \mathcal{O}_{iy}(K \times d)$
$\mathcal{N}(\mu_{iy}, \sigma_{iy})(2 \times K \times D)$

Estimer les Parametres:
Cas Bernoulli:

$$\mathcal{O}_{iy} = \frac{n(1, i, y)}{N(i, y)}$$

$$n(1, i, y) = \text{ nombre de fois où } X_i = 1 \text{ dans la classe } y$$

Si $n(1, i, y) = 0 \Rightarrow \mathcal{O}_{iy} = 0 \Rightarrow P(\vec{X} = \vec{x}|Y = y) = 0 \rightarrow$ mal
Estimation MLE (Maximum Likleyhood Estimate) $\rightarrow$ frequentiste

2 Inférence Bayesienne des paramètres (niveau 2)

On cherche $P(X_i|Y) \rightarrow P(X_i|Y, \underline{\mathcal{O}_{iy}})$ Apprendre le classifieur $\rightarrow$ estimer une distribution sur les paramètres

une variable aléatoire

Dans le 1):

~~MLE : $\mathcal{O}_{iy}$ maximise $P(\mathcal{D}|\mathcal{O}_{iy})$~~

Bayesien: Estimer $P(\mathcal{O}_{iy}|\mathcal{D})$

$$P(\mathcal{O}_{iy}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{O}_{iy})P(\mathcal{O}_{iy})}{P(\mathcal{D})}$$

ex: Bernoulli, $P(\mathcal{D}|\mathcal{O}_{iy}) \rightarrow$ facile

2.1 à priori sur les paramètres

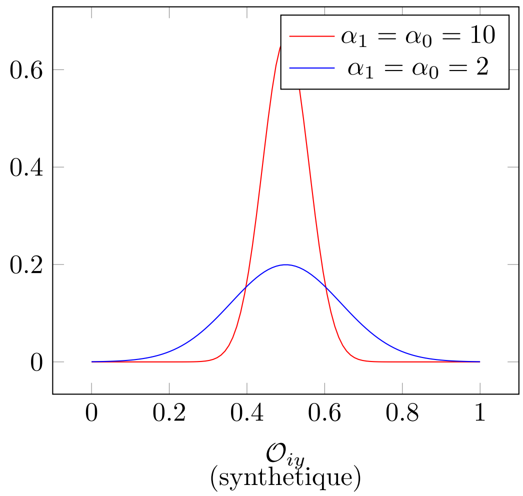
bernouilli: $\mathcal{O}_{iy} \in [0,1]$, continue $\rightarrow P(\mathcal{O}_{iy})$: une loi continue de support $[0,1]$

le choix: loi Beta

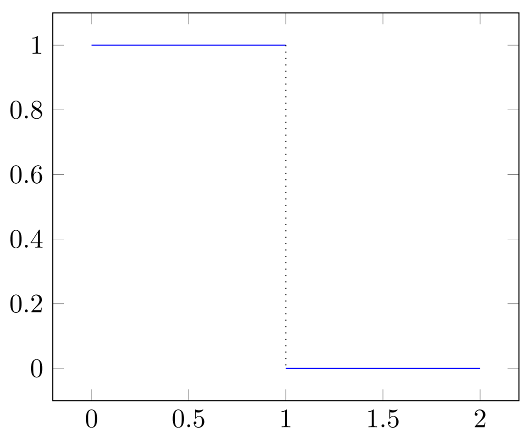
$$P(\mathcal{O}_{iy};\alpha_0,\alpha_1) = \underbrace{\frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}}_{\text{Normalisation}} \underbrace{\mathcal{O}_{iy}^{\alpha_1-1}(1-\mathcal{O}_{iy})^{\alpha_0-1}}$$

les parametres de la loi Beta $(\alpha_0, \alpha_1) > 0, \in \mathbb{R}$

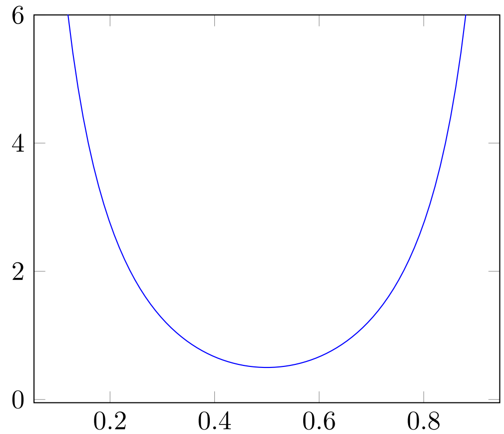
$\alpha_1 = \alpha_0 > 1$



•<u>A priori non-informatif</u> $(\alpha_1 = \alpha_0 = 1)$



•<u>A priori parcemonieux</u> (sparse)
$\alpha_1, \alpha_0 < 1$



2.2 à posteriori sur les paramètres

$$P(\mathcal{O}_{iy}|\mathcal{D}) \propto \underbrace{P(\mathcal{D}|\mathcal{O}_{iy})}\underbrace{P(\mathcal{O}_{iy};\alpha_1,\alpha_0)}_{\text{à priorie}} \propto \underline{\mathcal{O}_{iy}^{N_1+\alpha_1-1}(1-\mathcal{O}_{iy})^{N_0+\alpha_0-1}}$$

$N_1(N_0)$ nombre de x_i à 1(0) dans $\mathcal{D}$

La loi à posteriori est comme la loi à priori; une loi Beta.
La loi Beta est l'a priori <u>conjugé</u> de bernoulli. (Conjugated Prior)

2.3 Retour à la classification

a) <u>Maximum à Posteriri des paramètres (MAP)</u>
Dans le cas où α_0 et $\alpha_1 > 1$

$$\hat{\mathcal{O}_{iy}} = \underset{\mathcal{O}_{iy}}{\operatorname{argmax}} P(\mathcal{O}_{iy}|\mathcal{D}) = \frac{N_1 + \alpha_1 - 1}{N_1 + N_0 + \alpha_1 + \alpha_0 - 2}$$

$\mathcal{D}$(MLE) | à priori

•α_1 et α_0 agissent comme des <u>"pseudo-comptes"</u>→ lissage (smoothing) de distribution
•$\mathcal{O}_{iy} \neq 0$
•Si N_1,N_0 >> α_1,α_0 l'a priori negligable → Régularisation, écrit sur-apprentissage

b) <u>Loi predictive (inférence Bayesienne 3)</u>

$$P(X_i = x_i|Y = y; \mathcal{O}_{iy})$$

↰ estimer à partire de $\mathcal{D}$ (MAP)

La vraie prédiction:

$$P(X_i = x_i|\mathcal{D}) = \int_0^1 P(X_i = x_i, \mathcal{O}_{iy}|\mathcal{D}) d\mathcal{O}_{iy}$$

→ en marginalisant les paramètres.

$$\underbrace{P(X_i, \mathcal{O}_{iy}|\mathcal{D})} = \underbrace{P(X_i|\mathcal{O}_{iy}, \mathcal{D})}_{\text{vraisemblence}} \; \underbrace{P(\mathcal{O}_{iy}|\mathcal{D})}_{2.2)}$$

$$P(X_i = 1|\mathcal{D}) = \frac{N_1 + \alpha_1}{N_1 + N_0 + \alpha_1 + \alpha_0} \quad , \forall \alpha_1 \text{ et } \alpha_0 > 0$$