MAPSI — cours 8 : Regression logistique

Vincent Guigue
vincent.guigue@lip6.fr

LIP6 – Université Paris 6, France

Rappel sur les modèles génératifs

1. Choix d'une modélisation des données : $p(\mathbf{x}|\theta)$
2. Apprentissage = trouver θ
3. Application possible : décision bayesienne

$$r(\mathbf{x}) = \arg\max_k p(\theta_k|\mathbf{x}) = \frac{p(\mathbf{x}|\theta_k)p(\theta_k)}{p(\mathbf{x})}$$

4. Application bis : génération de $\tilde{\mathbf{x}} \sim \mathcal{D}(\theta_k)$

Apprentissage d'un modèle génératif $\Leftrightarrow$ Estimation de densité

- Estimer θ_k = estimer une densité de probabilité d'une classe k
- Hypothèse (forte) : les θ_k sont supposés indépendants
- Techniques d'estimation des θ_k

Maximum de vraisemblance

- $D_k = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ exemples supposés générés par $p(\mathbf{x}|\theta_k)$
  Seulement pour la classe k
- *Faire coller* le modèle au données

$$\mathcal{L}(D_k, \theta_k) = p(D_k|\theta_k) = \prod_{i=1}^{N} p(\mathbf{x}_i|\theta_k)$$

- Optimisation : $\theta_k^\star = \arg\max_{\theta_k}(\log \mathcal{L}(D_k, \theta_k))$
- Résolution :
  - Analytique : $\frac{\partial \mathcal{L}(D, \theta)}{\partial \theta} = 0$ ou Approchée : EM, gradient...
- Inférence sur une nouvelle donnée $\mathbf{x}$:

$$decision = k^\star = \arg\max_k p(\mathbf{x}|\theta_k)$$

Limite du modèle génératif pour la classification

- Approche **générative** : travail classe par classe

Quel modèle colle le mieux à mon observation ?

- Approche **discriminante** : travail classe i VS classe j

Qu'est ce qui distingue la classe i de la classe j ?

Idée : travailler sur les $p(Y|X)$
Modèle (le plus connu) : **Régression logistique**

Formulation du problème

- Echantillons $\{\mathbf{x}_i\}_{i=1,...,n}, \mathbf{x} \sim X$
- Deux classes : $Y = 0$ ou $Y = 1$. Réalisation des $y_i \sim Y$
- En fait : $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
- Comment modéliser $p(Y|X)$?

Formulation du problème

- Echantillons $\{\mathbf{x}_i\}_{i=1,\ldots,n}, \mathbf{x} \sim X$
- Deux classes : $Y = 0$ ou $Y = 1$. Réalisation des $y_i \sim Y$
- En fait : $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
- Comment modéliser $p(Y|X)$?

1. On repère une variable de Bernoulli
   $p(Y = 1|X = \mathbf{x}) = 1 - p(Y = 0|X = \mathbf{x})$
2. On choisit de poser arbitrairement :

$$p(Y = 1|X = \mathbf{x}) = f(\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{x}\mathbf{w} + b))}$$

- Inférence ?

Le problème devient :

Formulation du problème

- Echantillons $\{\mathbf{x}_i\}_{i=1,\ldots,n}, \mathbf{x} \sim X$
- Deux classes : $Y = 0$ ou $Y = 1$. Réalisation des $y_i \sim Y$
- En fait : $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
- Comment modéliser $p(Y|X)$?

1. On repère une variable de Bernoulli
   $p(Y = 1|X = \mathbf{x}) = 1 - p(Y = 0|X = \mathbf{x})$
2. On choisit de poser arbitrairement :

$$p(Y = 1|X = \mathbf{x}) = f(\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{x}\mathbf{w} + b))}$$

- Inférence ?

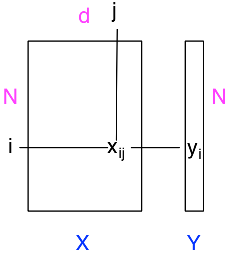
Le problème devient :

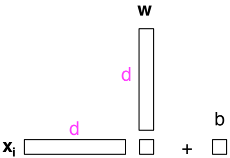
Trouver les $\mathbf{w}$ qui **distinguent** au mieux
la classe 1 de la classe 0

Dimension des éléments en présence

- Données : X, Y
- Modèle :
  $p(Y=1|X=\mathbf{x})=f(\mathbf{x})=\frac{1}{1+\exp(-(\mathbf{x}\mathbf{w}+b))}$
- Bornes du modèle :
  - $\lim\limits_{\mathbf{x}\mathbf{w}+b\to-\infty} f(\mathbf{x})=0$
  - $\lim\limits_{\mathbf{x}\mathbf{w}+b\to\infty} f(\mathbf{x})=1$

- Données : X, Y
- Modèle :
  $p(Y = 1|X = \mathbf{x}) = f(\mathbf{x}) = \frac{1}{1+\exp(-(\mathbf{xw}+b))}$
- Bornes du modèle :
  - $\lim_{\mathbf{xw}+b \to -\infty} f(\mathbf{x}) = 0$
  - $\lim_{\mathbf{xw}+b \to \infty} f(\mathbf{x}) = 1$
  - Si : $\mathbf{xw} + b = 0 \Rightarrow f(\mathbf{x}) = 0.5$





Comment trouver les $\mathbf{w}^\star$?

Comment trouver les $\mathbf{w}^\star$?

$\Rightarrow$ **Par maximum de vraisemblance (conditionnelle) !**
Vraisemblance (conditionnelle) -indépendance entre échantillons-

$$L = \Pi_{i=1}^{N} p(Y = y_i | X = \mathbf{x}_i)$$

- Truc de bernoulli :

$$p(Y = y_i | X = \mathbf{x}_i) = p(Y = 1 | X = \mathbf{x}_i)^{y_i} (1 - p(Y = 1 | X = \mathbf{x}_i))^{1-y_i}$$

- Passage au log
- Remplacement des $p(Y = 1|X = \mathbf{x})$ par des $f(\mathbf{x})$
- Nouvelle formulation :

$$\mathbf{w}^\star, b^\star = \arg\max_{\mathbf{w},b} \sum_{i=1}^{N} [y_i \log(f(\mathbf{x}_i)) + (1 - y_i)\log(1 - f(\mathbf{x}_i))]$$

Résolution

Données : $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

$$\mathbf{w}^\star, b^\star = \arg\max_{\mathbf{w},b} \sum_{i=1}^{N} [y_i \log(f(\mathbf{x}_i)) + (1 - y_i) \log(1 - f(\mathbf{x}_i))]$$

On remplace $f(x)$ par sa valeur et on développe le coût...

... [quelques lignes de calcul] ...

Résolution

Données : $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

$$\mathbf{w}^\star, b^\star = \arg\max_{\mathbf{w},b} \sum_{i=1}^{N} [y_i \log(f(\mathbf{x}_i)) + (1 - y_i) \log(1 - f(\mathbf{x}_i))]$$

On remplace $f(x)$ par sa valeur et on développe le coût...

... [quelques lignes de calcul] ...

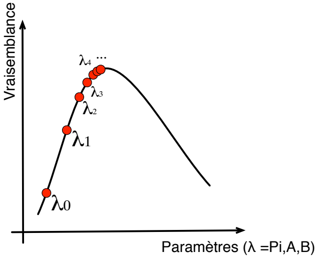
$$\frac{\partial}{\partial w_j} L_{\log} = \sum_{i=1}^{N} x_{ij} \left(y_i - \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))} \right)$$

$$\frac{\partial}{\partial b} L_{\log} = \sum_{i=1}^{N} \left(y_i - \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x}_i + b))} \right)$$

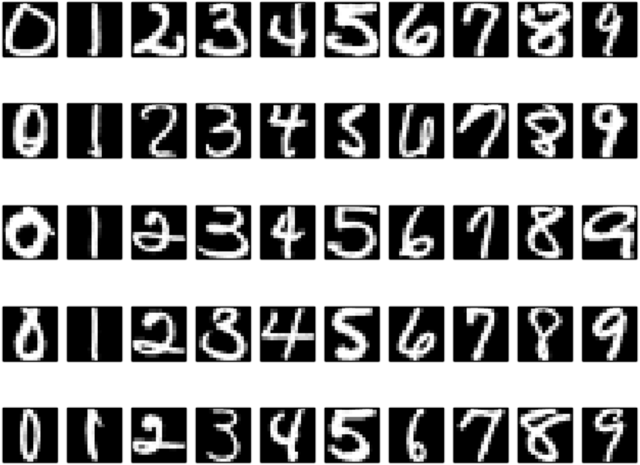
Résolution

- Annulation directe du gradient **impossible**
- $\Rightarrow$ Algorithme itératif :
  - Init des paramètres $\mathbf{w}_0, b_0$
  - Tant que convergence non atteinte
    - Calcul des : $\frac{\partial L_{\text{log}}}{\partial b}, \frac{\partial L_{\text{log}}}{\partial w_j}$
    - Mise à jour (montée de gradient) : $\theta^t = \theta^{t-1} + \epsilon \frac{\partial L_{\text{log}}}{\partial \theta}$

Cas convexe :



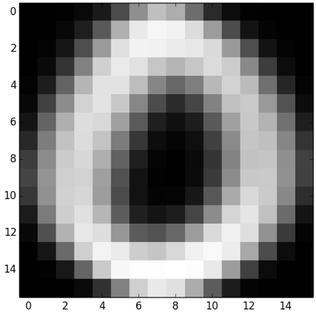
Type de résultats attendus



Type de résultats attendus

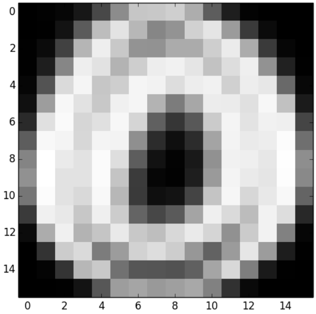
Modèle génératif gaussien : classe 0

Visualisation de la moyenne de la classe :



+ reshape

Visualisation de la variance de la classe :

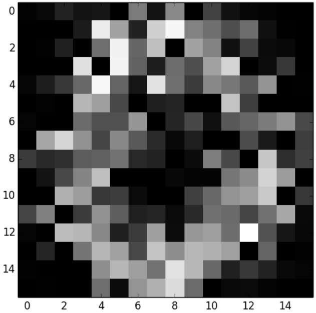


+ reshape

Type de résultats attendus

Possibilité de générer un échantillon :

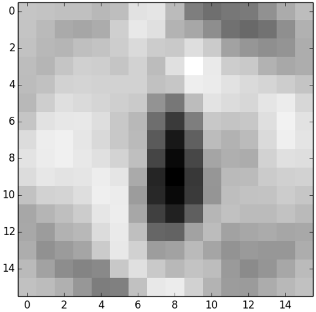
Tirage d'une valeur gaussienne pour chaque pixel...



Mais hypothèse d'indépendance des pixels $\Rightarrow$ qualité BOF

Type de résultats attendus

En régression logistique : classe 0 *VS* toutes les autres



Mise en évidence des zones qui ne sont utilisées **que** par les 0

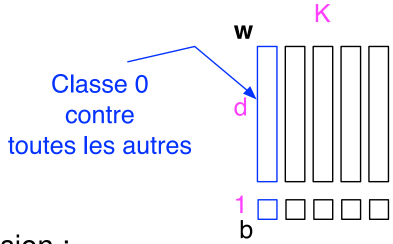
Type de résultats attendus

- Apprentissage : 7291 images
- Test : 2007 images

```
Naive Bayes (modèle de pixel gaussien)

Taux bonne classif. en apprentissage : 0.785
Taux bonne classif. en test : 0.739

Regression logistique

Taux bonne classif. en apprentissage : 0.943
Taux bonne classif. en test : 0.903
```

Comment passer au multi-classes ?

un contre tous (*one against all*) : K classes $\Rightarrow$ K classifieurs appris séparément sur **toutes les données**



- $f(\mathbf{x}) \Rightarrow f_k(\mathbf{x})$ et critère de décision :

$$k^\star = \arg\max_k f_k(\mathbf{x})$$

Quelle classe veut **le plus** l'échantillon **x** ?

- Critères de rejet :
  - pas de $f_k(\mathbf{x}) > 0.5$
  - plusieurs $f_k(\mathbf{x}) > 0.5$