Modélisation probabiliste

Vincent Guigue

Petit mémo de cours

A. Processus standard Bayésien :

1. **Récupération d'un jeu de données :**

$$D = \begin{bmatrix} x_1 & y_1 & z_1 \\ & \vdots & \\ x_\ell & y_\ell & z_\ell \\ & \vdots & \\ x_N & y_N & z_N \end{bmatrix}$$

Les individus sont sur les lignes, les variables explicatives sur les colonnes. L'échantillon est de taille N, décrit sur 3 variables. Les individus sont supposés iid : indépendant et identiquement distribués.

Hypothèse i.i.d. : [indépendance] les individus ne s'influencent pas les uns les autres ; [identiquement distribués] les individus viennent d'une distribution unique sous-jacente que nous allons chercher à caractériser.

2. **Choix d'une loi de probabilité pour modéliser tout ou partie des données :**

Vous êtes un expert et vous voulez modéliser la variable X. En fonction du type de X (discrète, continue), en fonction de ses modalités, en fonction de sa signification, vous allez choisir une loi de probabilité. Par exemple, la loi Normale (continue), Bernoulli (0/1), Binomiale, Géométrique, Exponentielle, ... Vous vous retrouvez alors avec une hypothèse de la forme :

$$p_\theta(X = x) = f(\theta, x), \qquad x : \text{observation}, \qquad \theta : \text{paramètre(s) de la loi de probabilité choisie}$$

3. **Optimisation des paramètres, maximum de vraisemblance :**

Trouver le(s) paramètre(s) θ optimal(aux) revient à faire coller notre modèle à nos observations. Pour cela, nous allons définir la vraisemblance, c'est à dire la probabilité d'observation de l'ensemble de l'échantillon X.

(a) Expression de la vraisemblance : $\mathcal{L}(X, \theta) = p(X|\theta) = p(X_1 = x_1, \ldots, X_\ell = x_\ell, \ldots, X_N = x_N)$

(b) Hypothèse i.i.d. : $\mathcal{L}(X, \theta) = \prod_\ell p(X_\ell = x_\ell)$. Dans la plupart des cas (modélisation avec des lois de probabilités classiques), la vraisemblance est convexe : elle n'admet qu'un maximum lorsque les dérivées en θ s'annulent.

(c) Maximisation de la vraisemblance = maximisation de la log-vraisemblance [1] : le log est une fonction croissante, plus la vraisemblance est grande, plus la log-vraisemblance est grande.

$$\theta^\star = \arg\max_\theta(\mathcal{L}(X, \theta)) = \arg\max_\theta(\log \mathcal{L}(X, \theta)), \qquad \text{même si bien sûr : } \mathcal{L}(X, \theta) \neq \log L(X, \theta)$$

(d) Opération de maximisation :

$$\nabla_\theta \mathcal{L}(X, \theta)|_{\theta^\star} = \begin{bmatrix} \frac{\partial \mathcal{L}(X,\theta)}{\partial \theta_1} \\ \frac{\partial \mathcal{L}(X,\theta)}{\partial \theta_2} \\ \vdots \end{bmatrix}_{\theta^\star} = 0$$

Lorsque nous avons obtenu $\theta^\star$, nous sommes en mesure de *traiter des données*, c'est à dire d'exploiter notre modèle en inférence... Mais comment et pour quoi faire ?

4. **Inférence :**

Pour n'importe quelle nouvelle entrée x, nous sommes en mesure de quantifier $p(X = x|\theta)$, c'est à dire à quel point le modèle colle à cette observation. Cela ouvre[irait] des perspectives en matière de détection d'anomalie en séparant les données vraisemblables des données aberrantes. Néanmoins, la définition de seuils est délicate et la plupart du temps, ces modèles sont plusieurs et nous comparons des valeurs $p(X = x|\theta^{(1)})$ par rapport à $p(X = x|\theta^{(2)})$.

L'étape 3 s'envisage sur feuille, il s'agit d'un travail analytique. L'étape 4 est a effectuer sur machine, c'est la phase d'implémentation.

1. Le log s'entend toujours au sens népérien, jamais en base 10.

5. **Maximum a posteriori :**
   Si un expert donne son avis sur les paramètres, sous la forme d'une distribution $p(\theta)$, il faut mêler l'optimisation des paramètres liés aux observations et à l'expert. Cela se fait très simplement, en maximisant $p(\theta|X)$

$$\arg\max_\theta p(\theta|X) = \arg\max_\theta \frac{\mathcal{L}(X,\theta)p(\theta)}{p(X)} = \arg\max_\theta(\mathcal{L}(X,\theta)p(\theta)), \qquad \text{car } p(X) > 0 \text{ ne dépend pas de } \theta$$

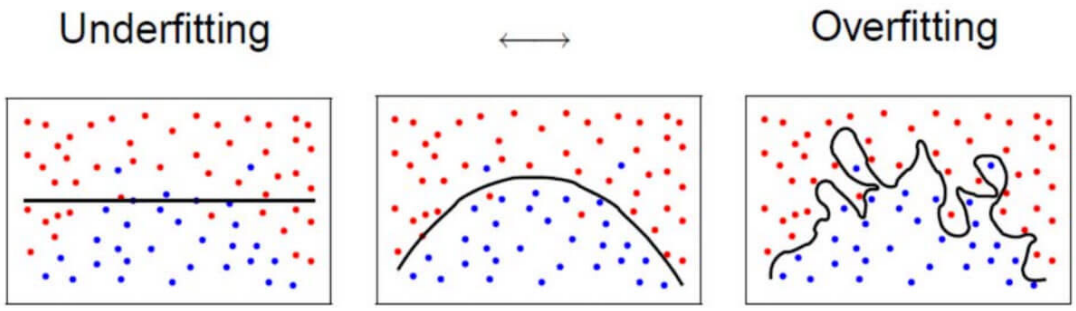
   L'optimisation est effectuée de la même manière que précedemment. De manière générale, cette approche est intéressante sur les petits jeux de données, où l'avis de l'expert permet d'éviter certains biais. Par contre, elle est souvent inutile sur les jeu de données plus conséquents.

B. Classification bayésienne supervisée : Dans le cadre défini précédemment, la modélisation d'un problème de classification s'entend classe par classe. Chaque classe est alors un univers indépendant.

1. **Données :**
   La notion d'étiquetage (en classes) implique une information supplémentaire par rapport à la présentation des données de la page précédente. De manière générale, soit il sera fourni un vecteur $Y = \{y_1, \dots, y_N\}$ contenant la classe associée aux données d'apprentissage, soit l'ensemble lui-même sera divisé en $D = \{D^{(1)}, \dots, D^{(C)}\}$ correspondant aux C classes.
2. **Apprentissage des modèles :**
   L'optimisation, indépendante, de C problèmes basés sur des données disjointes mène à un ensemble de paramètres : $\{\theta^{\star,1}, \dots, \theta^{\star,C}\}$. Ces paramètres sont issus de l'optimisation au sens du maximum de vraisemblance ou du maximum a posteriori, selon les données disponibles.
3. **Inférence :**
   Pour une donnée x, au sens du maximum de vraisemblance, il s'agit :

$$\text{de calculer : } out = \begin{bmatrix} p(X = x|\theta^{\star,1}) \\ \vdots \\ p(X = x|\theta^{\star,C}) \end{bmatrix}, \qquad \text{puis de prendre la décision : } \hat{y} = \arg\max_{index} out(index)$$

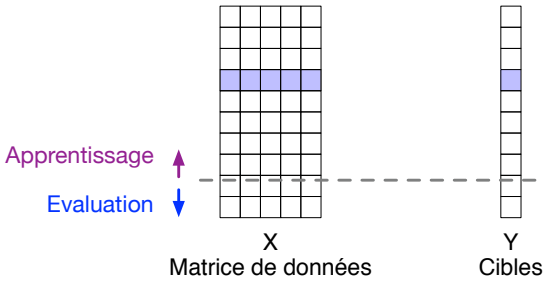
   Au sens du maximum a posteriori, la procédure est la même mais repose sur les $p(\theta^{\star,C}|X = x)$.
   Il est important de noter que nous pouvons indifféremment utiliser des probabilités ou des log-probabilités puisque la classe estimée ne dépend que de la position de la vraisemblance la plus forte (et pas de la valeur calculée de la vraisemblance).
   — Lors de l'apprentissage, nous passons au log pour facilité le calcul de la dérivée d'un produit (sur tous les individus de l'échantillon).
   — Lors de l'inférence, nous passons au log pour éviter les problèmes de précision machine qui peuvent survenir.
4. **Evaluation des modèles :**
   L'évaluation est une question critique en analyse de données : tout expert se doit de livrer un modèle en donnant sa performance ainsi que ses points faibles et ses points forts.
   L'évaluation des modèles n'est pas évidente. En effet, tester la performance de modèle sur les données qui ont servi à l'apprendre est une mauvaise idée entrainant des biais immédiats.



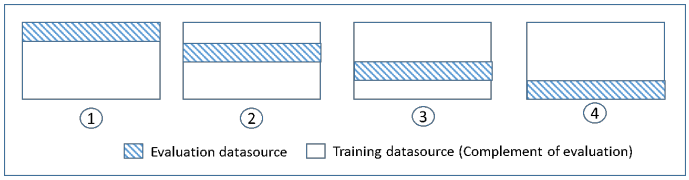
De ce point de vue, le meilleur modèle est celui de droite : bien que la solution ne soit pas satisfaisante, elle minimise bien l'erreur sur les données d'apprentissage.

Il est donc nécessaire d'utiliser des données vierges et étiquetées, données qui sont en générale chères et disponibles en quantité limitée :

— si nous utilisons trop de données en apprentissage, le modèle est bon mais l'évaluation faible.
— A l'inverse, une évaluation forte, sur beaucoup de données implique naturellement un modèle faible appris sur peu de données.



La solution consiste à prendre beaucoup de données en apprentissage et peu en évaluation, mais à itérer l'opération en apprenant *nval* modèles sur des sous-ensembles de données. A la fin du processus, notre modèle aura été évalué sur l'ensemble des données sans les avoir jamais vues en apprentissage. Cette procédure de validation croisée constitue une bonne estimation de l'erreur en généralisation.

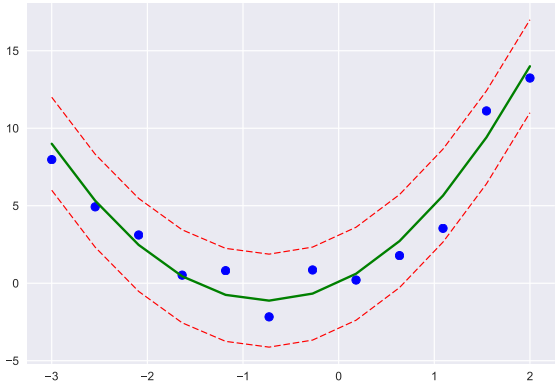


C. Régression : La modélisation bayesienne permet beaucoup de chose : il suffit de choisir le modèle qui nous semble raisonnable. Nous prenons volontairement un exemple assez différent de celui de la classification pour montrer les capacités générales du cadre défini précédemment.
Les points bleus représentent nos observations. Les courbes verte et rouge, ce que nous cherchons à apprendre : d'un part un modèle représentant une bonne estimation de l'ordonnée quelque soit l'abscisse, d'autre part, une estimation du niveau de bruit gaussien autour de ce modèle.

La figure suivante illustre un nuage de points (bleus), que nous souhaitons modéliser par un polynome de degré 2, en faisant l'hypothèse d'un bruit blanc gaussien d'écart type σ.

— Soit des observations : $\{(x_i, t_i), i = 1, \ldots, N\}$
— Nous notons $p(t|x, \mathbf{w}, \sigma^2) = \mathcal{N}(t|y(x, \mathbf{w}), \sigma^2)$, qui se lit : les données suivent une loi normale d'écart type σ centrée autour de $y(x, \mathbf{w})$, avec $y(x, \mathbf{w}) = w_0 + xw_1 + x^2w_2$

Rappel :
$A \sim \mathcal{N}(\mu, \sigma^2) \iff p(A = a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(a - \mu)^2\right)$



1. Vraisemblance de l'échantillon :

$$\mathcal{L} = \prod_{i=1}^{N} p(t_i|x_i, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(t_i - (w_0 + x_i w_1 + x_i^2 w_2))^2\right)$$

$$\log \mathcal{L} = \sum_i -0.5 \log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(t_i - (w_0 + x_i w_1 + x_i^2 w_2))^2$$

2. Optimisation au sens du maximum de vraisemblance :

$$\frac{\partial \log \mathcal{L}}{\partial w_0} = \sum_i \frac{1}{2\sigma^2}(-2)(t_i - (w_0 + x_i w_1 + x_i^2 w_2)) = 0 \iff N w_0 + (\sum_i x_i) w_1 + (\sum_i x_i^2) w_2 = \sum_i t_i$$

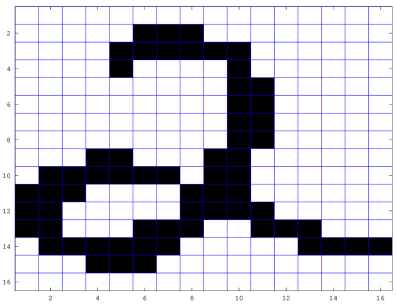
$$\frac{\partial \log \mathcal{L}}{\partial w_1} = 0 \iff \dots$$

$w_0^\star, w_1^\star, w_2^\star$ s'obtiennent par résolution d'un système de 3 équations à 3 inconnues puis :

$$\frac{\partial \log \mathcal{L}}{\partial \sigma} = \sum_i -\frac{1}{\sigma} + \frac{2}{2\sigma^3}(t_i - (w_0 + x_i w_1 + x_i^2 w_2))^2 = 0 \iff \sigma = \frac{1}{N}\sum_i (t_i - y_i)^2$$

3. En inférence, notre modèle est capable d'estimer la hauteur $\hat{t}$ de n'importe quel abscisse x.

Exercice 1 – Classification de données USPS : réflexions générales



Les images USPS sur lesquelles nous allons travailler sont des images de 16x16 pixels, par défaut en niveaux de gris. Nous pourrons les binariser ou pas en fonction des modèlisations choisies.

Nous nous intéressons à la modélisation de phénomène réels par des lois de probabilités standard.

Q 1.1 Dans une image $\mathbf{x}$, nous avons 256 pixels x_j noirs ou blancs. Quelle loi utiliser pour modéliser un pixel ? Que signifient le ou les paramètres de cette loi ?

Q 1.2 Imaginons que nous sommes dans un problème bi-classe, impliquant des chiens et des chats et que nous disposons de 2 modèles optimisés pour chaque classe. Nous ne nous intéressons qu'au pixel 18 dans un premier temps. Une nouvelle image arrive dont le pixel visible, x_{18}, est allumé : comment déterminer s'il s'agit d'un chien ou d'un chat ?

Q 1.3 Un expert nous indique l'importance des profils dans les images en noir et blanc : c'est à dire l'indice sur chaque ligne où se trouve le premier pixel allumé. Comment modéliser une ligne de l'image ? En imaginant que chaque ligne de l'image est indépendante, exprimer la probabilité d'observation de l'image $\mathbf{x}$ contenant 16 lignes $\{x_1, \dots, x_{16}\}$.

Q 1.4 Nous modélisation maintenant une image $\mathbf{x}$ dont les pixels x_j peuvent prendre 16 niveaux de gris différents. Quelle loi utiliser pour modéliser un pixel ? Que signifient le ou les paramètres de cette loi ?

Exercice 2 – Modélisation de Bernoulli des pixels

Il faut tout d'abord prendre en main les données et un notebook est disponible pour vous assister dans cette tâche. Une fois que vous avez saisi la structure des données, nous allons attaquer la modélisation.

Q 2.1 Soit une image binaire $\mathbf{x} \in \{0,1\}^{256}$, quelle est la loi de probabilité du pixel x_j ? Exprimer $p(X_j = 0)$, $p(X_j = 1)$, puis $p(X_j = x_j)$ dans le cas général.
Combien de paramètres sont nécessaires à la modélisation d'une classe d'image ?

Q 2.2 Exprimer la vraisemblance d'une image puis la vraisemblance de l'ensemble des images de la classe C.

Q 2.3 Optimiser les paramètres de la classe C

Q 2.4 Construire une chaine de traitement en python pour (1) apprendre les paramètres optimaux des classes de chiffres, (2) classer toutes les données de test, calculer un taux de bonne classification général et une matrice de confusion.

Exercice 3 – Modélisation Géométrique des profils de chiffres

Afin de reconnaitre des chiffres manuscrits (image binaire 16x16), on propose d'utiliser le modèle suivant :

- Chaque ligne j est décrite par une variable x_j donnant l'indice du premier pixel allumé sur la ligne j. x_j suit une loi géométrique de paramètre p_j : on part de la gauche, chaque pixel est une épreuve de Bernoulli et on attend le premier succès (pixel allumé).
- Afin de stabiliser les calculs, on ajoute une colonne (17) de pixels allumés.
- L'observation de chaque ligne est stockée dans une variable k_j donnant l'indice du premier pixel allumé. Par exemple sur l'illustration précédente : $k_1 = 17, k_2 = 6$...
- Les lignes sont supposées mutuellement indépendantes.

Notations : $\mathbf{x}$ est une image composée de N lignes : $\mathbf{x} = \{x_1, \dots, x_N\}$. On note $X = \{\mathbf{x}^1, \dots, \mathbf{x}^P\}$ un ensemble de P images, avec : $\mathbf{x}^i = \{x_1^i, \dots, x_N^i\}$.

Q 3.1 Quelles valeurs peuvent prendre les x_j ?
La loi géométrique correspond au fait de réussir une épreuve de Bernoulli au bout de k essais. En se posant sur la ligne j, en notant p_j le paramètre de la loi de Bernoulli pour cette ligne et k_j l'indice du premier pixel allumé de cette ligne, nous avons la formule suivante :

$$p(X_j = k_j) = p_j(1 - p_j)^{k_j - 1}$$

Q 3.2 Donner l'expression de la probabilité d'observation d'une image $p(\mathbf{x})$ si la valeur observée pour le pixel x_j est k_j. (Rappel : toutes les lignes sont indémendantes).

Q 3.3 Donner la log vraisemblance $\mathcal{L}$ de l'ensemble des images X, en supposant que toutes les images sont mutuellement indépendantes.

Q 3.4 Calculer les paramètres p_j maximisant la vraisemblance. Exprimer p_j en fonction de P et $K_j = \sum_i k_j^i$, où k_j^i représente la valeur observée du $j$ème pixel de la $i$ème image. Proposer une interprétation du résultat.

Q 3.5 Expliquer la procédure complète pour construire et évaluer un classifieur d'image au sens du maximum de vraisemblance basé sur ce modèle.

Q 3.6 On dispose maintenant d'un a priori $p(y)$ sur les différents chiffres (qui ne sont pas équiprobables dans la majorité des applications). Donner un critère de classification des images au sens du maximum à posteriori.

Exercice 4 – Modélisation en niveau de gris

Q 4.1 Passer les images en 16 niveaux de gris : soit en utilisant les histogrammes numpy, soit en repassant par `pandas` et la méthode `cut` qui fait ça parfaitement.

Q 4.2 Nous proposons d'utiliser une loi Binomiale. En notant x_j la valeur du pixel j (entre 0 et 15), et p_j le paramètre de la Bernoulli associée au pixel j, nous obtenons :

$$p(X_j = x_j) = \binom{16}{x_j} p_j^{x_j}(1 - p_j)^{16 - x_j}$$

Q 4.3 Nous pourrions aussi nous tourner vers la loi Multinomiale pour traiter ce cas de figure en posant directement pour chaque pixel j :

$$[p_{j,0} = p(X_j = 0), \dots, p_{j,15} = p(X_j = 15)]$$

Exercice 5 – Modélisation continue des pixels avec une loi normale

Q 5.1 Modéliser les pixels en utilisant la loi normale. Dans ce cas, les valeurs que peuvent prendre x_j sont des réels.

$$p(X_j = x_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{1}{2\sigma_j^2}(x_j - \mu_j)^2\right)$$

Q 5.2 Combien de paramètres sont à apprendre dans cette formulation ?

Q 5.3 La solution du maximum de vraisemblance est triviale pour la loi normale... Mais nous pouvons redémontrer ce résultat avec ceux qui le souhaitent.

Exercice 6 – Synthèse & validation croisée

Q 6.1 Comparer les sorties des différents modèles précédents : est-il possible de les combiner pour dépasser les résultats obtenus par le meilleur modèle unitaire ?

Q 6.2 Faire varier la taille de l'échantillon d'apprentissage pour vérifier l'importance d'avoir suffisamment de données en apprentissage.

Q 6.3 Implémenter une procédure de validation croisée comme défini dans le mémo en début d'énoncé.

Quelques exercices plus théoriques

Ces exercices ont vocation à servir d'illustration ou de support. Nous n'auront pas le temps de les traiter durant les TP, mais ils peuvent vous servir à mieux cerner les compétences que nous cherchons à vous transmettre.

Exercice 7 – MAP

Soit X une variable aléatoire définie sur l'ensemble des nombres entiers positifs. X suit la loi géométrique de paramètre $p \in [0,1]$ si $P(X = n) = (1-p)^{n-1}p$. On a observé 5 réalisations (obtenues indépendamment les unes des autres) d'une variable X suivant la loi géométrique :

4	2	6	5	8

.

Q 7.1 Estimez par maximum de vraisemblance la valeur du paramètre $\Theta = p$ de la loi.

Exercice 8 – Max de vraisemblance

Dans une urne se trouvent des boules de 4 couleurs différentes : rouge (R), bleues (B), vert (V) et jaune (J). On ne connaît pas la quantité de boules dans l'urne ni la proportion des différentes couleurs. Soit la variable aléatoire $X =$ « couleur d'une boule tirée au hasard dans l'urne ». On se propose de représenter la distribution de probabilité de X par une distribution catégorielle de paramètres $\theta = \{p_R, p_B, p_V, p_J\}$, c'est-à-dire :

$$P(X = R) = p_R \quad P(X = B) = p_B \quad P(X = V) = p_V \quad P(X = J) = p_J$$

avec, bien entendu, $p_R, p_B, p_V, p_J \geq 0$ et $p_R + p_B + p_V + p_J = 1$.
Afin d'estimer les paramètres de la distribution, on a tiré avec remise un échantillon des boules de l'urne et on a observé leurs couleurs, que l'on a retranscrites dans le tableau suivant :

R	R	R	R	B	B	V	V	V	J

Q 8.1 Estimez par maximum de vraisemblance les paramètres de la distribution $P(X)$. Vous justifierez votre réponse.

Exercice 9 – Maximum a posteriori, maximum de vraisemblance

Une pièce de monnaie peut être plus ou moins biaisée en faveur de *Pile* ou de *Face*.
On prend pour paramètre θ la probabilité de *Pile* :

$$P_\theta(Pile) = \theta.$$

L'ensemble des valeurs possibles pour θ est $\Theta = \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$; les probabilités a priori $\pi(\theta)$ de la v.a. $\tilde{\theta}$ sont :

θ	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$
$\pi(\theta)$	0.1	0.2	0.4	0.2	0.1

On effectue 5 lancers indépendants de la pièce et on observe le nombre x de résultats *Pile* obtenus ; la v.a. X a donc pour valeurs possibles $\mathcal{X} = \{0, 1, 2, 3, 4, 5\}$.

Q 9.1 Quelle est la loi suivie par X conditionnellement à l'hypothèse $\tilde{\theta} = \theta$? Calculer tous les éléments du tableau des probabilités conditionnelles $P(x|\theta)$, $(x, \theta) \in \mathcal{X} \times \Theta$. (on pourra se servir d'une table et mettre à profit les symétries des données)

Q 9.2 Déduire de la question précédente les valeurs des éléments du tableau des probabilités jointes $\pi(x,\theta),\ (x,\theta) \in \mathcal{X} \times \Theta$. À partir de ce tableau, comment peut-on retrouver la loi a priori $\{\pi(\theta)\}$ de la v.a. $\tilde{\theta}$? comment trouve-t-on la loi a priori de X ? Calculez-la.

Q 9.3 Déduire de ce qui précède les valeurs des éléments du tableau des probabilités a posteriori $\pi(\theta|x),\ (x,\theta) \in \mathcal{X} \times \Theta$.

Q 9.4 Donner les valeurs d'acceptation des diverses hypothèses sur la valeur du paramètre :

Q 9.4.1 quand la règle de décision est celle de la *probabilité d'erreur minimum* ; Cette règle équivaut à une règle de la probabilité maximum d'une décision juste : dans chaque ligne du tableau des $\pi(\theta|x)$ il faut choisir

$$d(x) = \arg\max_{\theta} \pi(\theta|x).$$

Q 9.4.2 quand la règle de décision est celle du *maximum de vraisemblance*.

Q 9.4.3 Quand ces deux règles donnent-elles le même résultat ?

Exercice 10 – Loi exponentielle et MAP

La loi exponentielle est une loi continue dont la fonction de densité est : $f(x) = \lambda e^{-\lambda x}$ pour tout $x > 0$. Elle sert, entre autres, pour caractériser la durée de vie des composants électroniques. Le tableau suivant recense les durées de vie (en années) observées pour un échantillon de 10 composants électroniques :

2	7	3	4	1	2	6	5	1	9

Q 10.1 On suppose que la distribution des durées de vie est effectivement une loi exponentielle. Estimez par maximum de vraisemblance la valeur de λ.

Q 10.2 Après discussion avec un expert en électronique, on a un a priori sous la forme d'une loi Gamma de densité $g(x) = \frac{1}{\Gamma(5)} x^4 e^{-x}$. Estimez par maximum a posteriori la valeur de λ.

Exercice 11 – Loi géométrique et maximum de vraisemblance

Un robot effectue des actions et, afin de déterminer son efficacité, un observateur a noté les temps d'exécution (en secondes) de 100 tâches qu'il a effectuées. Ces temps sont indiqués dans le tableau ci-dessous :

temps (en secondes)	1	2	3	4	5	6	7	8	9
nb observations	31	22	15	11	7	5	4	2	3

Q 11.1 L'observateur pense que la variable aléatoire X=« temps d'exécution » suit une loi géométrique. On rappelle que la loi géométrique de paramètre p est telle que $P(X = k) = p(1-p)^{k-1}$, pour tout entier $k \geq 0$. Déterminez la valeur du paramètre p par maximum de vraisemblance.