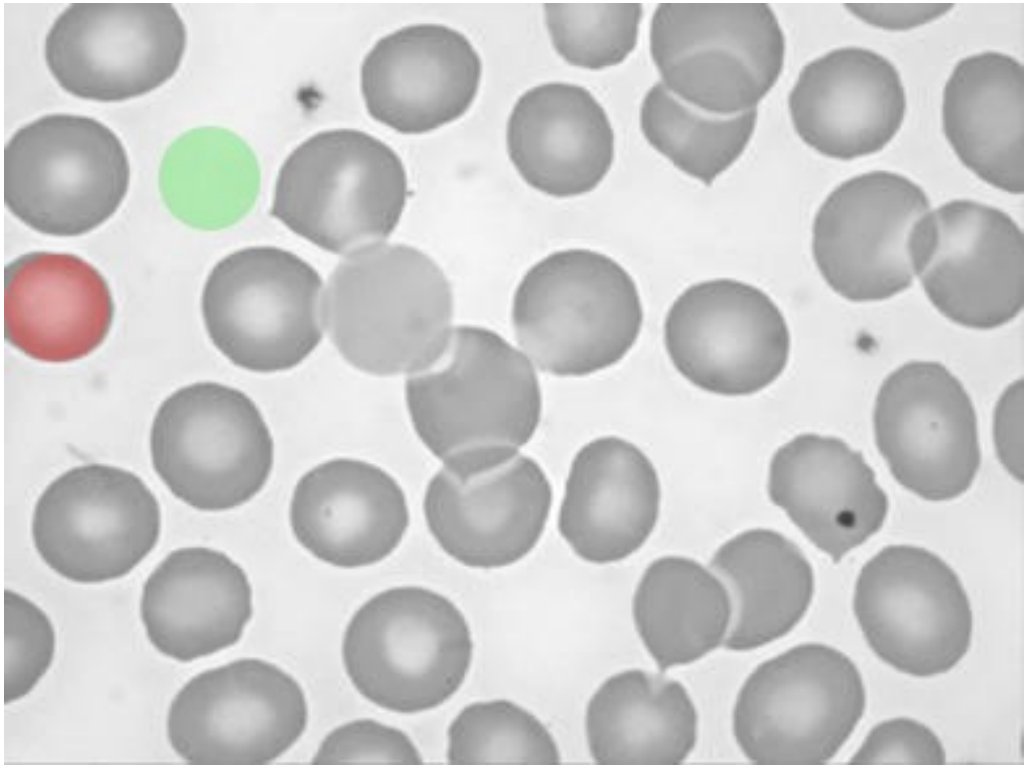
ESTIMATION

TMI

Exemple

Problème

- Image de cellules : à séparer du fond
- Deux classes : cellule et fond
- Annotation partielle
- En déduire une annotation totale ?



Exemple

Approche : minimiser le risque d'erreur de classification

- Trouver la meilleure classe C ∈ {C_1=cellule, C_2=fond} pour chaque valeur x de pixels
- Formule de Bayes $$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

- $P(C_k|x)$: probabilité a posteriori de la classe C_k, connaissant l'observation x
- $P(x|C_k)$: vraisemblance de l'observation x, étant donnée la classe C_k
- $P(C_k)$: probabilité a priori de la classe C_k
- P(x) : probabilité a priori de l'observation x

Exemple

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

Considérations et hypothèses

- $P(C_k|x)$ est à maximiser de manière globale pour chaque x → correspond à minimiser le risque d'erreur de classification
- P(x) ne va pas changer : les observations sont ce qu'elles sont

$$P(C_k|x) \propto P(x|C_k)P(C_k)$$

- $P(C_k)$: (ici) pas de connaissance a priori sur la fréquence relative des classes donc on suppose que c'est uniforme

$$P(C_k|x) \propto P(x|C_k)$$

- Maximiser la probabilité a posteriori équivaut à maximiser la vraisemblance

Exemple

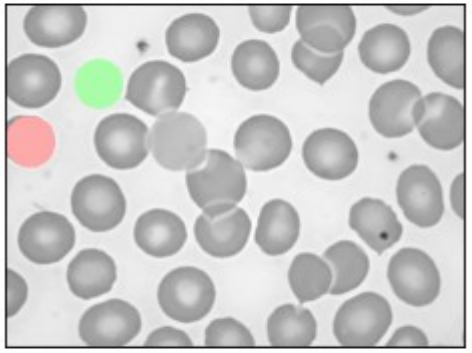
Liens avec l'estimation

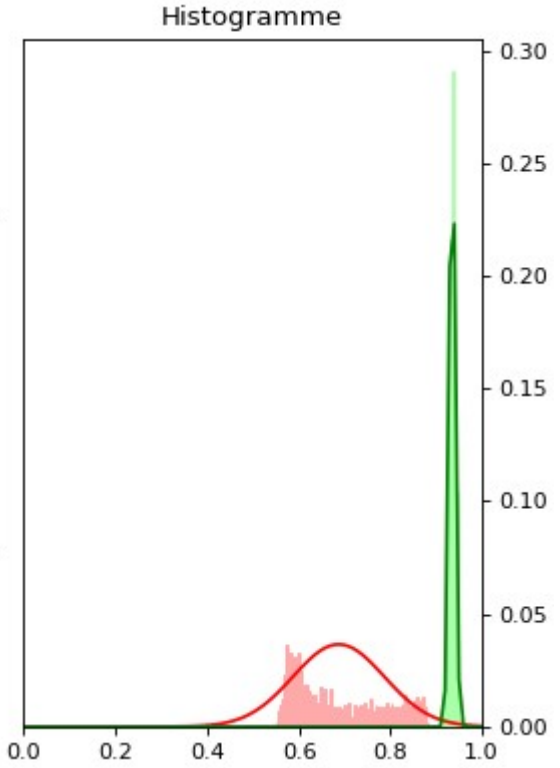
$$P(C_k|x) \propto P(x|C_k)$$

- Vraisemblance → modéliser la génération des observations aléatoires en fonction de la classe
- Modèle de loi de probabilité pour chaque classe → paramètres de la loi
- Ex : Lois gaussiennes → (μ_c,σ_c) ; (μ_b,σ_b)

Objectif

- Estimer ces paramètres, à partir des observations





PLAN DU COURS

Rappels d’estimation statistique

- Définition et propriétés d’un estimateur
- Exemple fondamental : estimateurs de la moyenne et de la variance
- Paramètres d’une loi : maximum de vraisemblance
- Qualité d’un estimateur

ESTIMATEUR : DÉFINITION

Contexte

Nous disposons de plusieurs mesures d'un même phénomène (*observations*)

$x=(x_1,x_2,\ldots,x_n)$

Ex : lancers de la même pièce, mesures de la température d'un patient, valeurs des pixels d'une image, paires de points mis en correspondance

Ces mesures sont une réalisation d'une variable aléatoire $X=(X_1,X_2,...X_n)$, tels que les X_i sont de même loi et indépendantes (i.i.d.) : *n-échantillon*

Ce phénomène est caractérisé par une grandeur θ (peut être un vecteur)

Objectif

Déterminer une valeur approchée θ^* de la vraie valeur θ_0

Estimateur $\theta^* = g(X) = g(X_1,X_2,\ldots X_n)$

ESTIMATEUR : PROPRIÉTÉS

Risque

Le risque quadratique est $R(\theta,\theta_0) = E[(\theta-\theta_0)^2]$

Un estimateur θ_1 est meilleur qu'un autre estimateur θ_2 si

$$\forall \theta_0, R(\theta_1,\theta_0) \leq R(\theta_2,\theta_0) \text{ et } \exists \theta_0, R(\theta_1,\theta_0) < R(\theta_2,\theta_0)$$

Biais

Le biais de l'estimateur est $b(\theta,\theta_0) = E(\theta)-\theta_0$.

Si le biais est nul, l'estimateur est dit sans biais ou non biaisé

Décomposition biais-variance et compromis

On peut montrer que $R(\theta,\theta_0) = b(\theta,\theta_0)^2+Var(\theta)$

Critère de variance minimale pour des estimateurs sans biais

EXEMPLE FONDAMENTAL : ESTIMATEUR DE LA MOYENNE

On considère que les X_i suivent tous une loi d'espérance m et de variance s^2

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Variable aléatoire fonction des X_i → estimateur

Calcul du biais

$$E[\bar{X}] = E[\frac{X_1 + X_2 + \ldots + X_n}{n}] = \frac{E[X_1] + E[X_2] + \ldots + E[X_n]}{n}$$

or $\forall i, E[X_i] = m$

donc $E[\bar{X}] = \frac{nm}{n} = m$

Son biais est $b(\bar{X}, m) = E[\bar{X}] - m = 0$

EXEMPLE FONDAMENTAL : ESTIMATEUR DE LA MOYENNE

$$Var(\bar{X}) = E[(\bar{X} - E[\bar{X}])^2] = E[\left((\frac{1}{n}\sum_{i=1}^{n} X_i) - m\right)^2]$$

$$Var(\bar{X}) = E[\left(\frac{1}{n}(\sum_{i=1}^{n} X_i - nm)\right)^2] = \frac{1}{n^2} E[\left(\sum_{i=1}^{n} (X_i - m)\right)^2]$$

$$Var(\bar{X}) = \frac{1}{n^2}\left(\sum_{i=1}^{n} E[(X_i - m)^2] + \sum_{i \neq j} E[(X_i - m)(X_j - m)]\right)$$

or $\forall i, j \; E[(X_i - m)(X_j - m)] = 0$ (indépendance, implique non corrélation)

donc $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} E[(X_i - m)^2]$

or $\forall i, E[(X_i - m)^2] = s^2$ (même loi)

donc $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} s^2 = \frac{n s^2}{n^2} = \frac{s^2}{n}$

La variance de cet estimateur est asymptotiquement nulle

EXEMPLE FONDAMENTAL : ESTIMATEUR DE LA VARIANCE

Cas où m est connu

Estimateur $S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2$

Sans biais $E[S^2]=\frac{1}{n}\sum_{i=1}^{n}E[(X_i-m)^2]=s^2$

Et on peut montrer que sa variance tend vers 0

Cas où m est inconnu

Estimateur $S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\bar{X}^2$

Asymptotiquement sans biais $E[S^2]=\frac{1}{n}\sum_{i=1}^{n}E[X_i^2]-E[\bar{X}^2]$

$$E[S^2]=\frac{1}{n}\sum_{i=1}^{n}(m^2+s^2)-\left(m^2+\frac{s^2}{n}\right)=\frac{n-1}{n}s^2$$

Et on peut aussi montrer que sa variance tend vers 0

Exercice : montrer que $\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ est un estimateur sans biais de S^2

Maximum de vraisemblance

Principe

Trouver le (un) paramètre θ qui maximise la probabilité d'observer les mesures dont on dispose : $P(X \mid \theta)$

Quand cette probabilité est vue comme une fonction de θ, on parle de fonction de vraisemblance.

Énonce du problème

Déterminer : $\theta^* = \text{argmax}_\theta P(X \mid \theta) = \text{argmax}_\theta g_X(\theta)$

θ^* est l'estimateur au maximum de vraisemblance du paramètre θ

EXEMPLE : CAS DISCRET

Énoncé

10 lancers d'une pièce ont donné : F,F,F,P,F,F,P,P,F,F (=X)

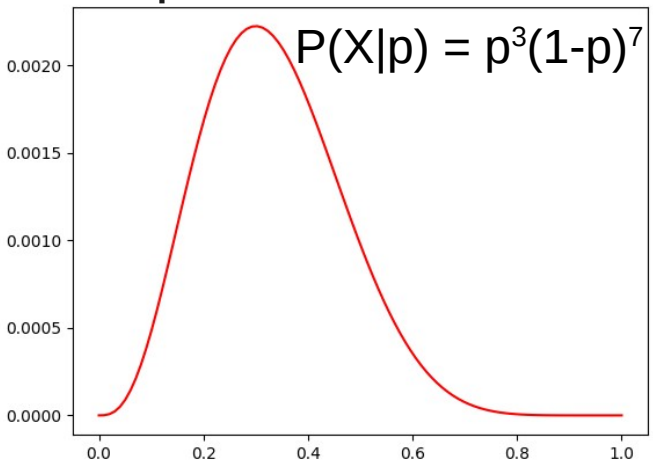
Quelle est la loi que suivent ces lancers ?

Résolution

Hypothèse : loi binomiale, de paramètre p=probabilité de tomber sur P

Calculs de la vraisemblance pour diverses valeurs de p

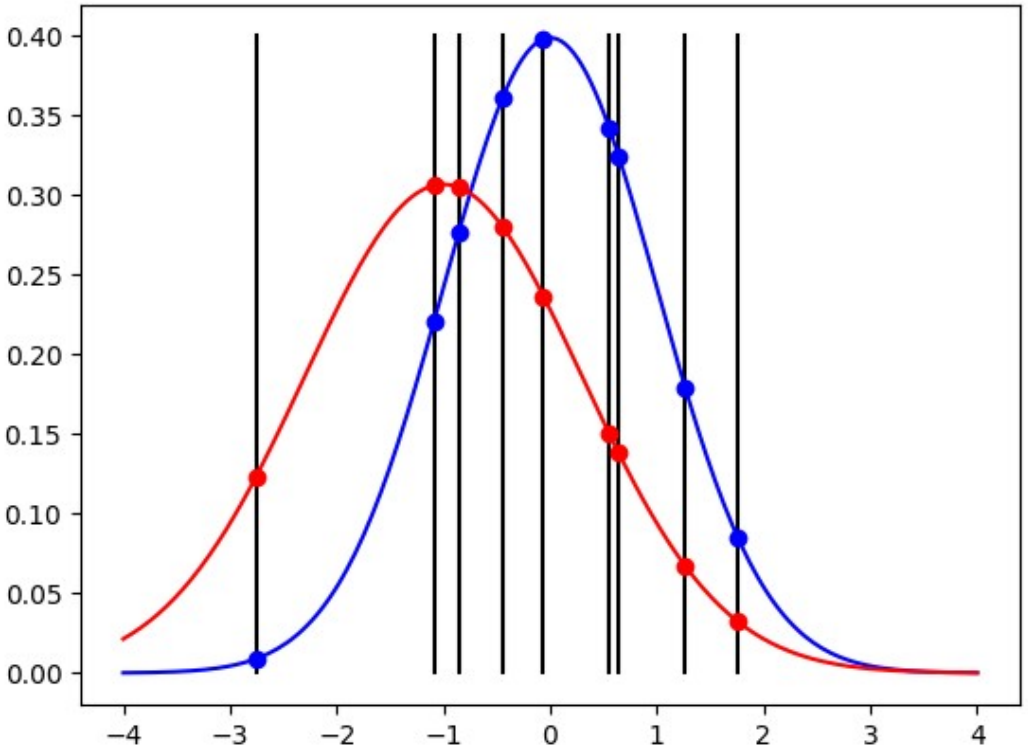
- p=0.5 : $P(X|p=0.5)=0.5^3*0.5^7=0.000976$
- p=0.25: $P(X|p=0.25)=0.25^3*0.75^7=0.00208$
- p=0.75 : $P(X|p=0.75)=0.75^3*0.25^7=0.0000257$



EXEMPLE CONTINU (TIRÉ DE WIKIPEDIA)

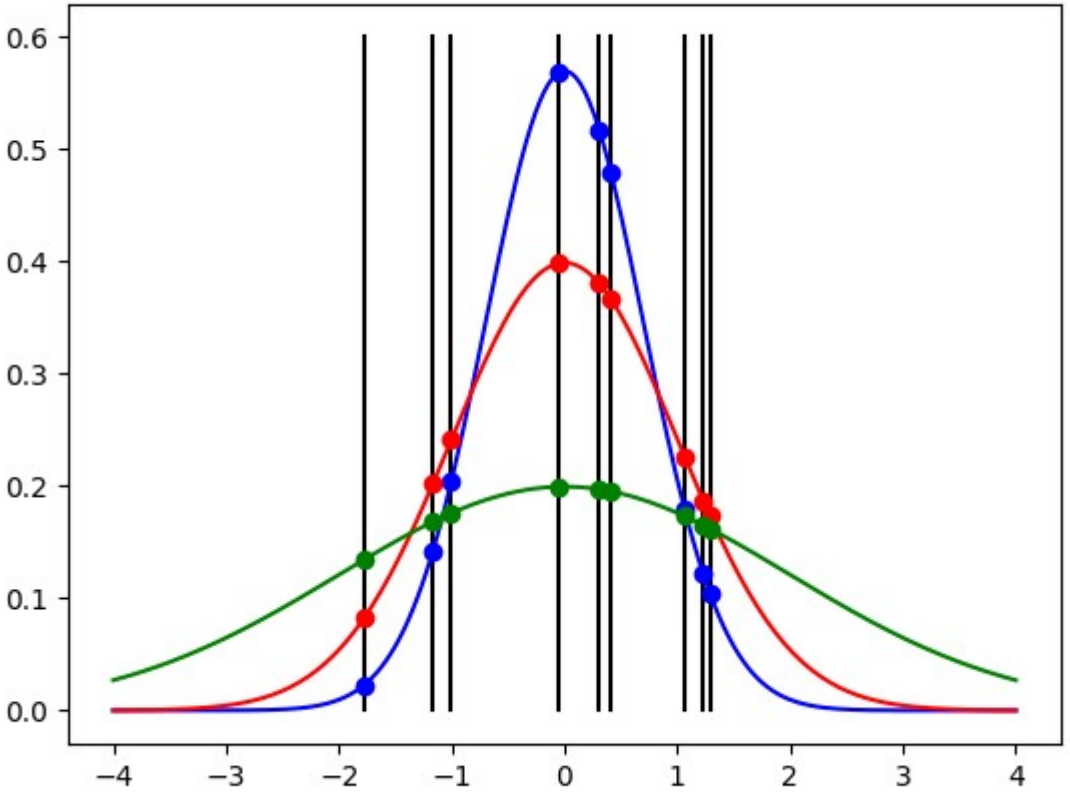
Soient 9 tirages aléatoires $x_1,\ldots,x_9$ suivant chacun une même loi gaussienne (traits verticaux). On souhaite modéliser ces valeurs par une loi normale. Parmi les deux courbes rouge et bleue, laquelle fournit la meilleure estimation du maximum de vraisemblance ?

Vraisemblance = produit des ordonnées des points

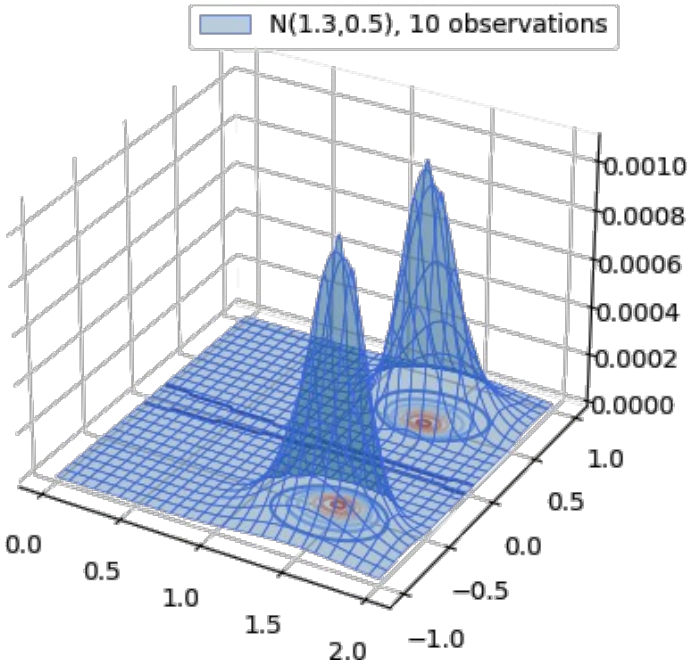


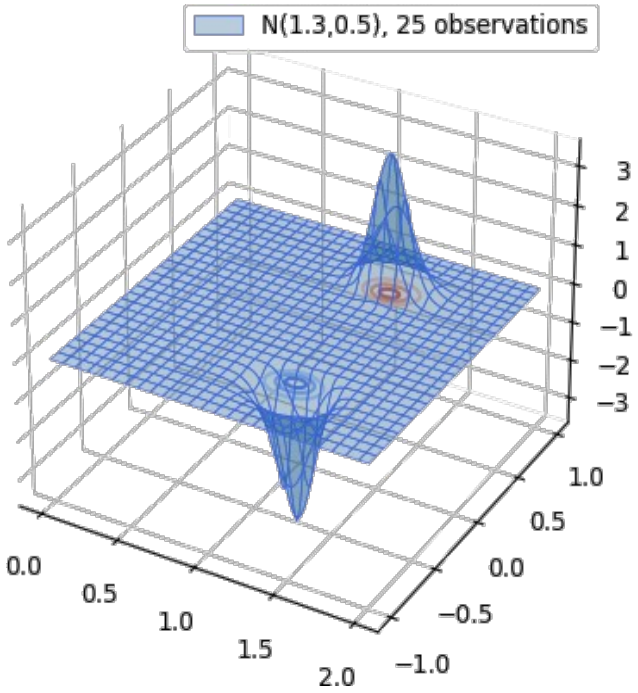
EXEMPLE CONTINU (TIRÉ DE WIKIPEDIA)

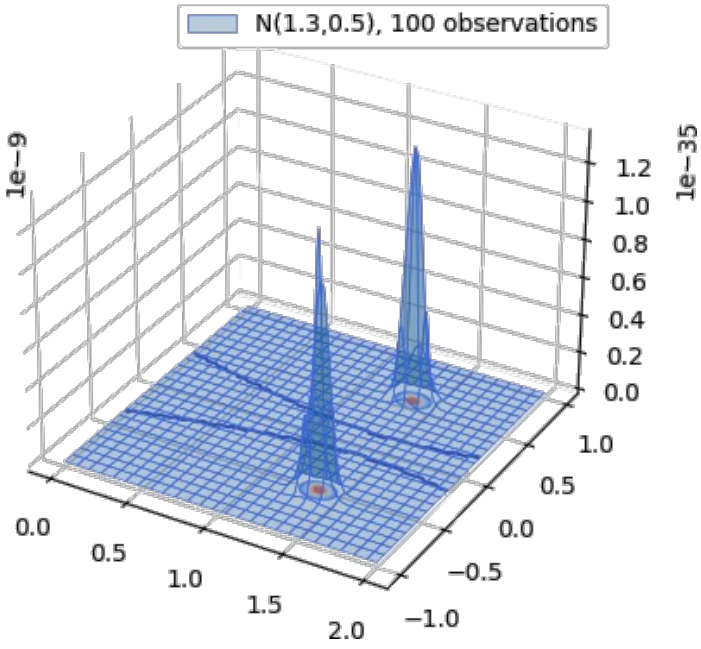
Et dans ce cas ?



VRAISEMBLANCE D'UNE VARIABLE GAUSSIENNE







Méthode explicite de calcul

Hypothèses

Un n-échantillon $X=(X_1,...X_n)$. Tous les X_i sont i.i.d. de loi f paramétrée par θ

Méthode explicite

Comme les X_i sont indépendantes, on peut écrire :

$$P(X|\theta)=\prod_{i=1}^{n} f(X_i|\theta)$$

Soit, en prenant le log (fonction croissante) :

$$\log(P(X|\theta))=\sum_{i=1}^{n} \log(f(X_i|\theta))$$

Si f est dérivable (densité), une condition nécessaire est :

$$\sum_{i=1}^{n} \frac{\partial \log(f(X_i|\theta))}{\partial \theta}=0$$

Exemple

Estimation de la moyenne d'une distribution gaussienne

Observations $x=(x_1,...x_n)$ suivant une loi gaussienne d'écart-type σ connu et de moyenne μ inconnue.

Quelle est l'estimateur de μ au maximum de vraisemblance ?

Loi: $P(X|\mu)=\frac{1}{\sqrt{2\pi}\sigma}\exp\frac{-(X-\mu)^2}{2\sigma^2}$

Log (naturel): $\ln(P(X|\mu))=-\ln(\sqrt{2\pi}\sigma)-\frac{(X-\mu)^2}{2\sigma^2}$

Dérivée : $\frac{\partial \ln(P(X|\mu))}{\partial \mu}=\frac{2(X-\mu)}{2\sigma^2}=\frac{X-\mu}{\sigma^2}$

Condition : $\sum_{i=1}^{n}\frac{\partial \ln(P(x_i|\mu))}{\partial \mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0\Leftrightarrow\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}\mu\Leftrightarrow\mu=\frac{1}{n}\sum_{i=1}^{n}x_i$

INFORMATION DE FISCHER

Qualité de l'estimation

Au maximum, la fonction de log vraisemblance a une courbure négative

Information de Fischer

$$I(\theta) = -\frac{\partial^2 \log(P(X|\theta))}{\partial \theta^2}$$

Exemple : loi gaussienne

Dérivée première: $\sum_{i=1}^{n} \frac{\partial \ln(P(x_i|\mu))}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$

Dérivée seconde: $I(\mu) = -\sum_{i=1}^{n} \frac{\partial^2 \log(P(x_i|\mu))}{\partial \mu^2} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} -1 = \frac{n}{\sigma^2}$

EXEMPLE : APPRENTISSAGE NON SUPERV

Exemple tiré de [DH73]

- Ensemble de mesures, issues de 2 classes
- Classe inconnue mais on connaît $p(C_1)=1/3$ et $p(C_2)=2/3$
- Probabilités gaussiennes : $p(x|C_1)$ et $p(x|C_2)$ de variance 1 et de moyennes respectivement μ_1 et μ_2 inconnues

Problème

- Estimer μ_1 et μ_2 à partir des observations (x_i)
- $p(x|\mu_1, \mu_2)$?

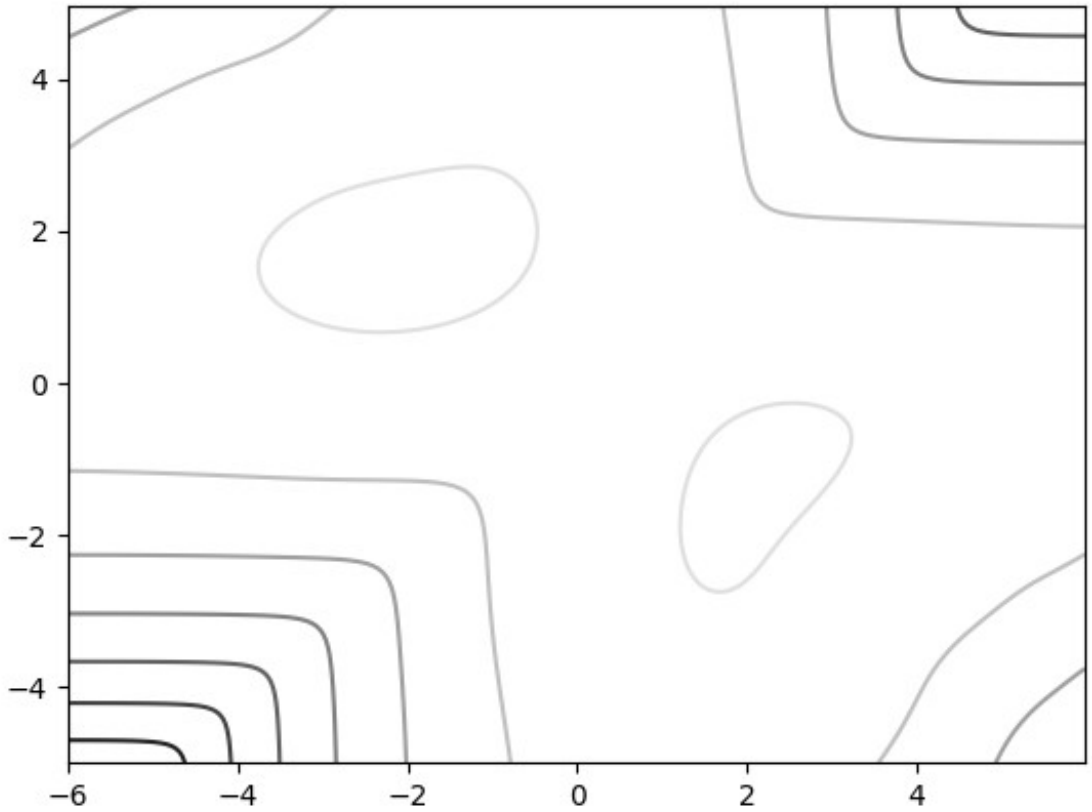
$$p(x|\mu_1,\mu_2)=p(x|C_1)\,p(C_1)+p(x|C_2)\,p(C_2)$$

$$=\frac{1}{3\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_1)^2}{2}\right)+\frac{2}{3\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_2)^2}{2}\right)$$

Index	x_i	Index	x_i
1	0.608	14	2.400
2	-1.590	15	-2.499
3	0.235	16	2.608
4	3.949	17	-3.458
5	-2.249	18	0.257
6	2.704	19	2.659
7	-2.473	20	1.415
8	0.672	21	1.410
9	0.262	22	-2.653
10	1.072	23	1.396
11	-1.773	24	3.286
12	0.537	25	-0.712
13	3.240		

[DH73] R.O. Duda and P.E. Hart. *Pattern recognition and Scene Analysis.* Wiley-In

EXEMPLE : LOG VRAISEMBLANCE



Deux maxima pour (μ_1, μ_2) : (-2.130,1.668) et (2.085,-1.257)

Nécessité de recourir à des méthodes numériques

CAS LINÉAIRE

Modèle linéaire $Y = w^T X + Z$

Paires d'observations (x_i, y_i) telles que

- x_i sont des vecteurs et y_i des réels
- w est le paramètre d'une transformation linéaire (vecteur), à estimer
- Z est un bruit blanc gaussien d'écart-type σ, connu

Maximum de vraisemblance

Loi: $P(X, Y | w, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(Y - w^T X)^2}{2\sigma^2}$

Log (naturel): $\ln(P(X, Y | w, \sigma)) = -\ln(\sqrt{2\pi}\sigma) - \frac{(Y - w^T X)^2}{2\sigma^2}$

Équation à résoudre: $\sum_{i=1}^{n} (y_i - w^t x_i)^2 = 0 \quad \Rightarrow \quad$ Moindres carrés

EXTENSION MULTIDIMENSIONNELLE

Cas général multidimensionnel

On peut étendre à la combinaison linéaire de fonction de X

Modèle : $Y=\sum_{k=1}^{p}\theta_k\,\phi_k(X)+Z$

On dispose de n mesures (x_i,y_i), Z bruit blanc d'écart type σ

Critère : $\sum_{i=1}^{n}\left(y_i-w_i^t\,\theta\right)^2$ où $w_i=\left[\phi_k(x_i)\right]_{1\le k\le p}$

Réécriture : $(y-A\,\theta)^t(y-A\,\theta)=y^t\,y-\theta^t\,A^t\,y-y\,A\,\theta+\theta\,A^t\,A\,\theta$

Dérivée : $2\,A^t\,A\,\theta-2\,A^t\,y$

Annulation pour : $\theta=\left(A^t\,A\right)^{-1}A^t\,y$

Solution de : $y=A\,\theta$ pour A non carrée

Extension si mesures covariantes : $\theta=\left(A^t\,\Sigma^{-1}\,A\right)^{-1}A^t\,\Sigma^{-1}\,y$

Reste valable si bruit pas gaussien : minimisation de la distance de Mahalanobis

MOINDRES CARRÉS NON LINÉAIRES

Relation non linéaire

$$y_i = f_i(\theta) \Rightarrow \hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f_i(\theta))^2$$

Dans ce cas, pas de solution directe (en général) : usage de minimisation numérique pour trouver θ

QUALITÉ D'ESTIMATION

Cas général

$\theta^* = \text{argmin}_\theta C(\theta)$ (C=log vraisemblance ou critère adapté au problème)

Développement de Taylor (ordre 2)

$$C(\theta) \approx C(\theta_0) + \nabla C(\theta_0)^t \delta\theta + \frac{1}{2} \delta\theta^t H \delta\theta \quad \text{où} \quad \delta\theta = (\theta - \theta_0)$$

Où H est la matrice hessienne de C en θ_0

Comme au maximum le gradient est nul :

$$C(\theta) - C(\theta_0) \approx \frac{1}{2} \delta\theta^t H \delta\theta$$

Un estimé est acceptable si $\delta\theta$ tel que $C(\theta) - C(\theta_0) < \epsilon$

Soit $\delta\theta^t H \delta\theta < 2\epsilon$

Ce qui forme une ellipsoïde de dimension dim(θ)

Qualité d'un estimateur : Ensemble de test

Principe

- Garder une partie des données qui ne participent pas à l'estimation
- Évaluer les métriques (par exemple critère à optimiser) sur ces données
- Les données qui participent à l'évaluation sont les données d'entraînement

Problèmes

- Cela nécessite beaucoup de données
- Compliqué dans un contexte de développement du modèle → ajouter ensemble de validation pour fixer les hyperparamètres

QUALITÉ D'UN ESTIMATEUR : VALIDATION CROISÉE

Principe (à k plis, voir exemple sur notebook)

On sépare l'ensemble des données en k sous-ensembles → k plis

On considère k estimations : k-1 plis pour l'estimation, 1 pli pour le test

On obtient k valeurs pour la métrique de qualité → statistiques possibles

Variantes

Leave-p-out : On va considérer tous les sous-ensembles possibles de p données comme ensemble de test et utiliser le reste pour l'estimation.

Très vite gourmand

Variante fréquente : leave-one-out

Intermédiaire : Monte-Carlo

Critique

La variance sur l'estimé du critère est sur-évaluée en général

EXEMPLE INITIAL : ALGORITHME EM

Calcul paramètres de classe

μ_c=0.687, σ_c=0.09989

μ_b=0.935, σ_b=0.00645

Résultat classification

Recalcul

μ_c=0.742, σ_c=0.11994

μ_b=0.929, σ_b=0.01144

Nouveau résultat

Algorithme EM

Peut s'initialiser sans label