Processus Gaussien

Cours 8
ML Master DAC

Nicolas Baskiotis
nicolas.baskiotis@lip6.fr
http://webia.lip6.fr/~baskiotisn

équipe MLIA, Laboratoire d'Informatique de Paris 6 (LIP6)
Sorbonne Université

S2 (2020-2021)

Plan

Régression et noyaux

Formulation

Pour un jeu de données $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^N$ $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$ i.i.d.

- On se donne un noyau $K(\mathbf{x}^1, \mathbf{x}^2) = \langle \phi(\mathbf{x}^1), \phi(\mathbf{x}^2) \rangle$, avec $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$
- Régression pénalisée : $J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^t \phi(\mathbf{x}^i) - y^i)^2 + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w}$
- On note $\Phi \in \mathbb{R}^{N \times d}$ la matrice des $\phi(\mathbf{x}^i)$,

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^t \Phi^t - \mathbf{y}^t)(\mathbf{w}^t \Phi^t - \mathbf{y}^t)^t + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w} = \frac{1}{2} \mathbf{w}^t \Phi^t \Phi \mathbf{w} - \mathbf{w}^t \Phi^t \mathbf{y} + \frac{1}{2} \mathbf{y}^t \mathbf{y} + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w}$$

Régression et noyaux

Annulation du gradient

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^t\Phi^t - \mathbf{y}^t)(\mathbf{w}^t\Phi^t - \mathbf{y}^t)^t + \frac{\lambda}{2}\mathbf{w}^t\mathbf{w} = \frac{1}{2}\mathbf{w}^t\Phi^t\Phi\mathbf{w} - \mathbf{w}^t\Phi^t\mathbf{y} + \frac{1}{2}\mathbf{y}^t\mathbf{y} + \frac{\lambda}{2}\mathbf{w}^t\mathbf{w}$$

- Annulation du gradient par rapport à $\mathbf{w}$:

$$\mathbf{w} = -\frac{1}{\lambda}\sum_{i=1}^{N}(\mathbf{w}^t\phi(\mathbf{x}^i) - y^i)\phi(\mathbf{x}^i) = \sum_{i=1}^{N} a^i\phi(\mathbf{x}^i) = \Phi^t\mathbf{a}$$

  avec $\mathbf{a} = (a^1, \ldots, a^N)$
- En ré-écrivant, avec $\mathbf{w} = \Phi^t\mathbf{a}$:

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^t\Phi\Phi^t\Phi^t\Phi\mathbf{a} - \mathbf{a}^t\Phi\Phi^t\mathbf{y} + \frac{1}{2}\mathbf{y}^t\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^t\Phi\Phi^t\mathbf{a}$$

Régression et noyaux

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^t\Phi\Phi^t\Phi^t\Phi\mathbf{a} - \mathbf{a}^t\Phi\Phi^t\mathbf{y} + \frac{1}{2}\mathbf{y}^t\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^t\Phi\Phi^t\mathbf{a}$$

Résolution

- On note $K = \Phi^T\Phi$, avec $K_{i,j} = \langle\phi(\mathbf{x}^i), \phi(\mathbf{x}^j)\rangle = k(\mathbf{x}^i, \mathbf{x}^j)$, la matrice de Gram du noyau (symétrique), on a

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^tK^tK\mathbf{a} - \mathbf{a}^tK^t\mathbf{y} + \frac{1}{2}\mathbf{y}^t\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^tK^t\mathbf{a}$$

- En prenant le gradient par rapport à $\mathbf{a}$, on trouve

$$\mathbf{a} = (K + \lambda I_N)^{-1}\mathbf{y}$$

- Pour la prédiction : $y = \mathbf{w}^t\phi(\mathbf{x}) = \mathbf{a}^t\Phi\phi(\mathbf{x}) = \mathbf{k}(\mathbf{x})^t(K + \lambda I_N)^{-1}\mathbf{y}$, avec $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}^1, \mathbf{x}), \ldots, k(\mathbf{x}^N, \mathbf{x}))^t$

Hypothèse du bruit gaussien

Formalisation

- Un jeu de données $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^N$ $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$ i.i.d.
- On suppose que $y^i = f(\mathbf{x}^i) + \epsilon^i$
- Régression linéaire : f de la forme $\mathbf{w}^t.\mathbf{x}$ (on oublie le biais pour simplifier)
- Le petit "plus" par rapport à la régression linéaire "simple" : on suppose $\epsilon^i \sim \mathcal{N}(0, \sigma^2)$, indépendant de $\mathbf{x}^i$

$\Rightarrow$ la distribution de y conditionnée au modèle et à l'entrée $\mathbf{x}$ est gaussienne :

$$p(y^i|\mathbf{x}^i; \mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^i - \mathbf{w}^t.\mathbf{x}^i)^2}{2\sigma^2}}$$

Hypothèse du bruit gaussien

Résolution par maximum de vraisemblance

- $L(\mathbf{w}, \sigma^2) = \prod_{i=1}^{N} p(y^i|\mathbf{x}^i; \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^i - \mathbf{w}^t.\mathbf{x}^i)^2}{2\sigma^2}}$
- $logL(\mathbf{w}, \sigma^2) = -\frac{N}{2} log 2\pi \ - N log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^i - \mathbf{w}^t.\mathbf{x}^i)^2$

$\rightarrow$ gradient par rapport à $\mathbf{w}$: $\mathbf{w} = (X^T X)^{-1} X^t \mathbf{y}$

$\rightarrow$ gradient par rapport à σ : $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y^i - \mathbf{w}^t.x^i)^2$

Quelle est l"avantage alors de rajouter l'hypothèse de bruit gaussien ?

$\Rightarrow$ On connaît pour une entrée $\mathbf{x}$ la distribution de l'estimée $\hat{y}$ qui suit une loi gaussienne . . .

Plan

Une gaussienne et tout est gaussien !

Soit $\mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$, $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $p(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(\mathbf{y}-\mu)^t \Sigma^{-1}(\mathbf{y}-\mu)}$

On considère une partition en 2 groupes : $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

Alors

- $\Sigma_{12}^t = \Sigma_{21}$
- la somme de deux gaussiennes est gaussienne : $\mathbf{y}_1 + \mathbf{y}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma_{11} + \Sigma_{22})$
- la marginalisation est gaussienne :

$$p(\mathbf{y}_1) = \int_{\mathbf{y}_2} p(\mathbf{y}_1, \mathbf{y}_2; \boldsymbol{\mu}, \Sigma) d\mathbf{y}_2 = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$$

- la conditionnée est gaussienne :

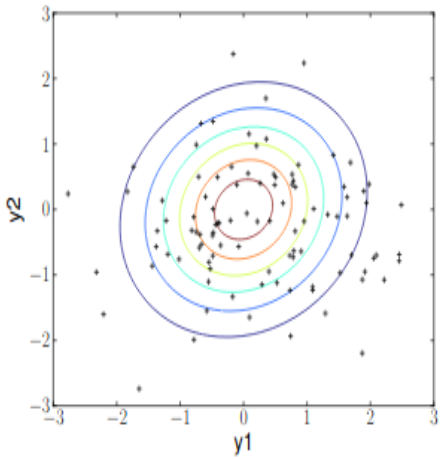
$$p(\mathbf{y}_2 | \mathbf{y}_1 = \mathbf{a}; \boldsymbol{\mu}, \Sigma) = \frac{p(\mathbf{y}; \mu, \Sigma)}{\int_{\mathbf{y}_2} p(\mathbf{y}; \mu, \Sigma) d\mathbf{y}_2}$$

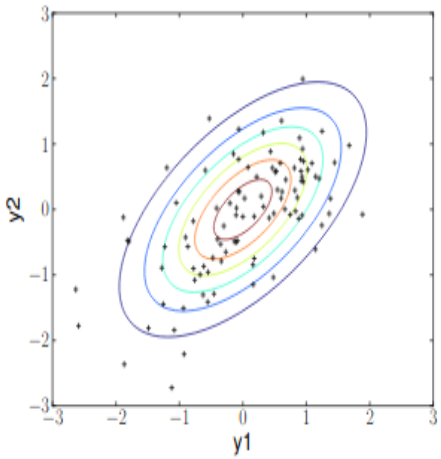
et suit la loi

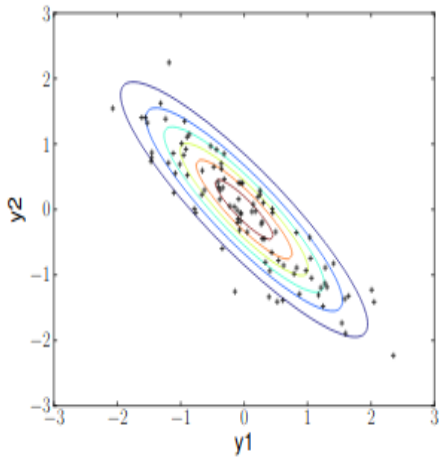
$$\mathcal{N}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{a} - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{12}^t \Sigma_{22}^{-1} \Sigma_{12})$$

Effet de la covariance

$$p(\mathbf{y}_2|\mathbf{y}_1 = \mathbf{a}; \boldsymbol{\mu}, \Sigma) \sim \mathcal{N}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{a} - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{12}^{t}\Sigma_{22}^{-1}\Sigma_{12})$$



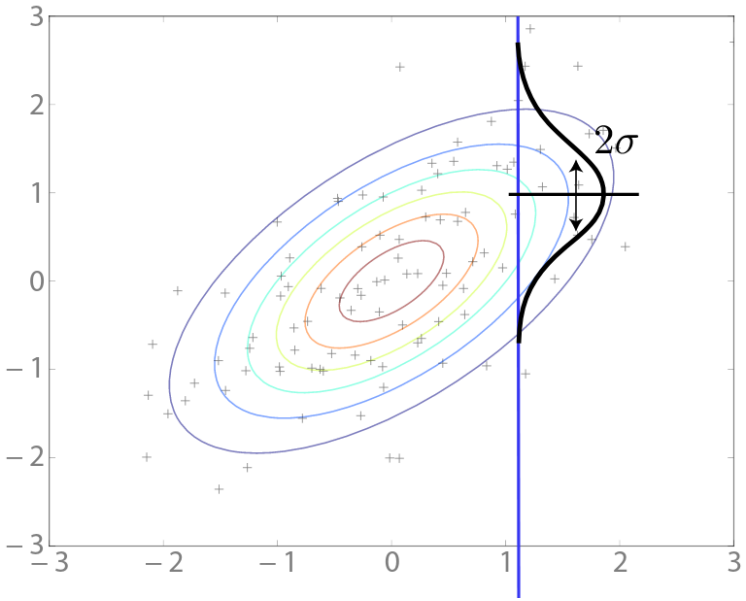


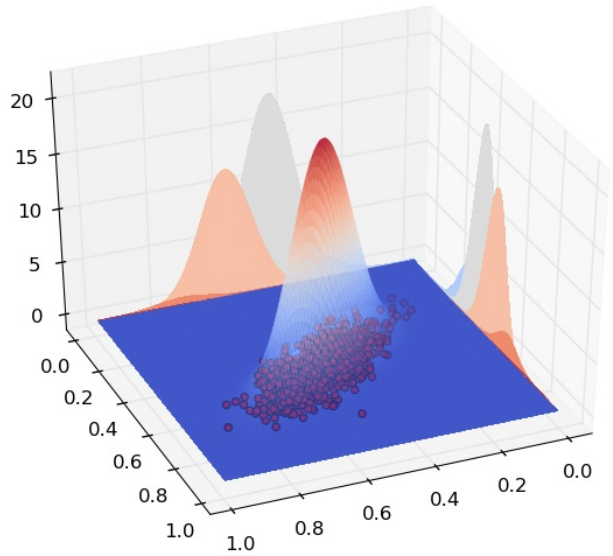


$$\Sigma = \begin{bmatrix} 1 & 0.14 \\ 0.14 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

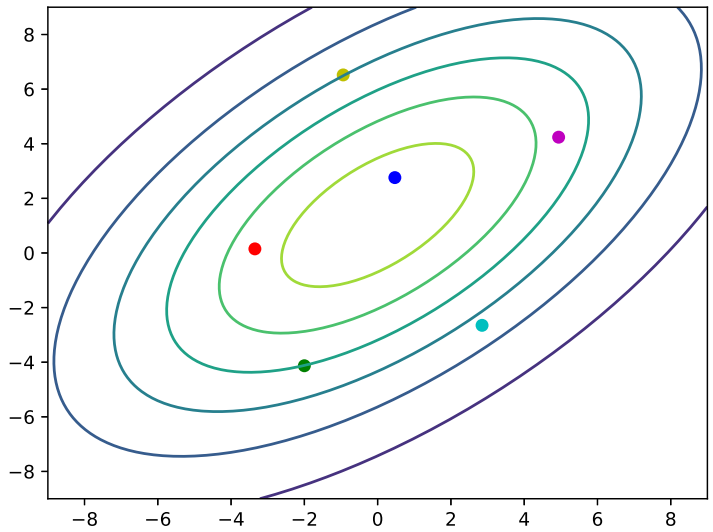
Corrélation entre coordonnées

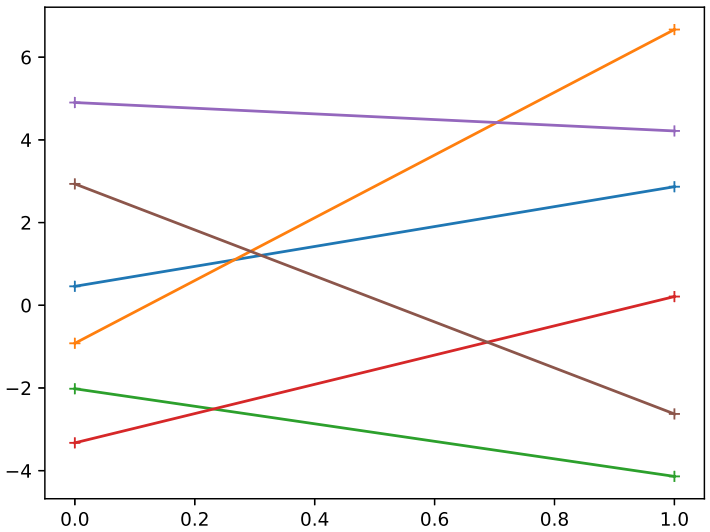
$$p(\mathbf{y}_2|\mathbf{y}_1 = \mathbf{a}; \boldsymbol{\mu}, \Sigma) \sim \mathcal{N}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{a} - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{12}^{t}\Sigma_{22}^{-1}\Sigma_{12})$$





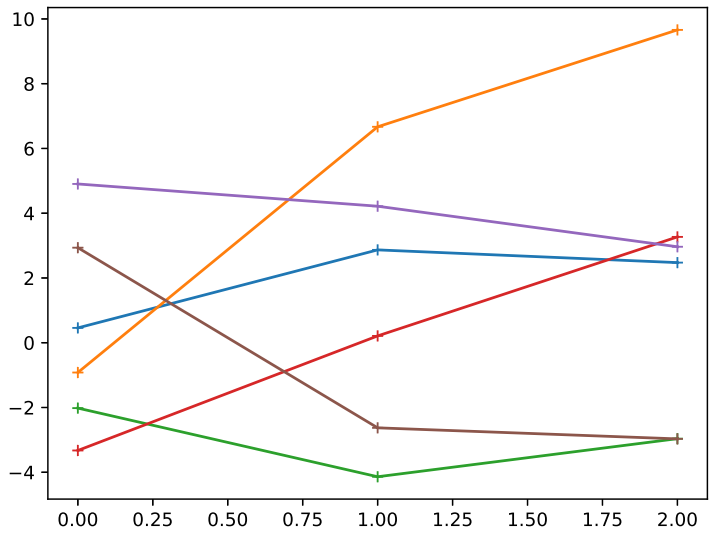
Une autre manière de visualiser une gaussienne

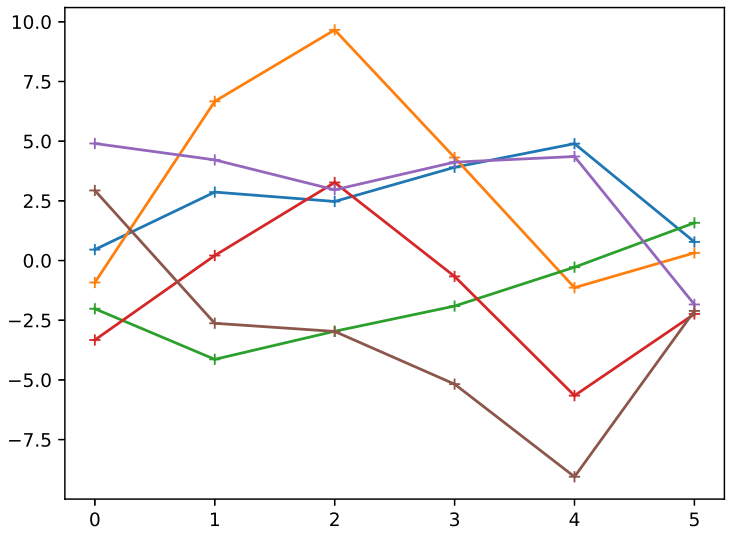




Pour chaque point 2D tirée de cette gaussienne, la première coordonnée est placée en 0, sa deuxième en 1.

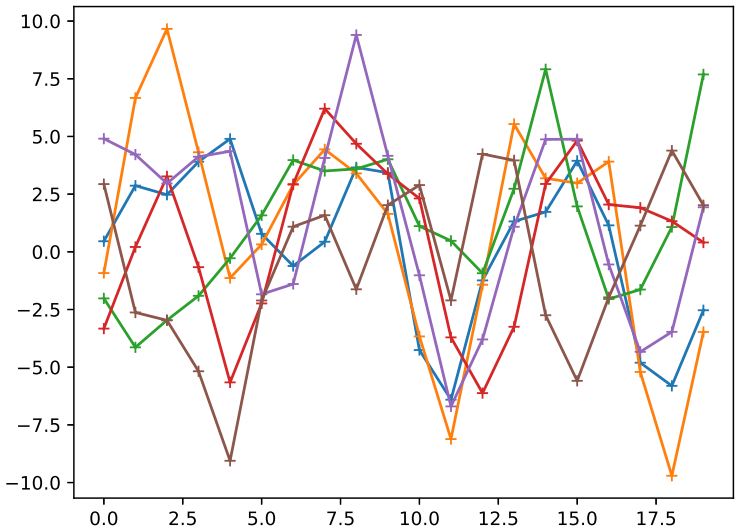
En 3d et plus . . .

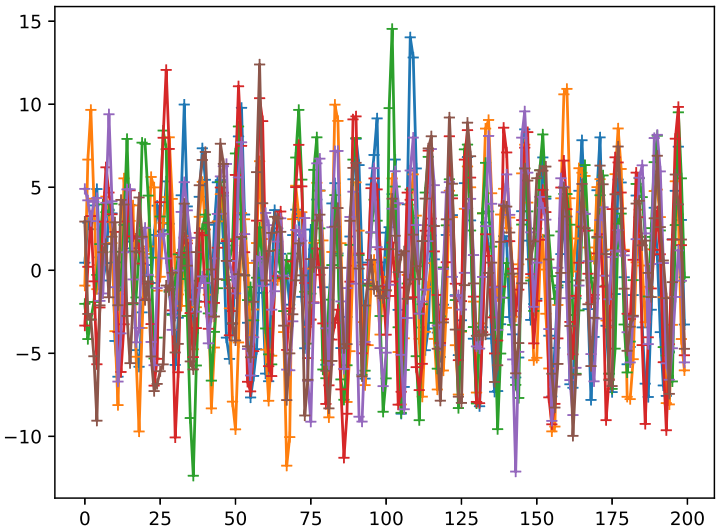




Soit une gaussienne en N dimensions, la i-ème coordonnée est placée en $x = i$.
(rappel : la relation entre la i-ème et j-ème dimension est définie par la covariance Σ_{ij})

En 3d et plus . . .





Soit une gaussienne en N dimensions, la i-ème coordonnée est placée en $x = i$.
(rappel : la relation entre la i-ème et j-ème dimension est définie par la covariance Σ_{ij})

Plan

Régression linéaire bayésienne

Rappel : $f(\mathbf{x}) = \mathbf{w}^t \phi(\mathbf{x})$ (déterministe) et $y = f(\mathbf{x}) + \epsilon$ (avec ϵ bruit gaussien) et des données $D = \{\mathbf{x}^i, y^i\}_{i=1}^N$
Processus habituel : de D on estime $\mathbf{w}$, puis on fait les prédictions.

Où sont les gaussiennes ?

- $p(y|\mathbf{x}; \mathbf{w})$: gaussien
- La vraisemblance : $p(D|\mathbf{w}) = \prod_{i=1}^N p(y^i|\mathbf{x}^i; \mathbf{w}) \Rightarrow$ gaussien
- Le prior : $p(\mathbf{w})$ est gaussien (dans le cadre de la ridge régression)
- Le posterior : $p(\mathbf{w}|D) = \frac{p(\mathbf{w})p(D|\mathbf{w})}{p(D)} \Rightarrow$ gaussien

Régression linéaire bayésienne

Rappel : $f(\mathbf{x}) = \mathbf{w}^t \phi(\mathbf{x})$ (déterministe) et $y = f(\mathbf{x}) + \epsilon$ (avec ϵ bruit gaussien) et des données $D = \{\mathbf{x}^i, y^i\}_{i=1}^N$
Processus habituel : de D on estime $\mathbf{w}$, puis on fait les prédictions.

Où sont les gaussiennes ?

- $p(y|\mathbf{x}; \mathbf{w})$: gaussien
- La vraisemblance : $p(D|\mathbf{w}) = \prod_{i=1}^N p(y^i|\mathbf{x}^i; \mathbf{w}) \Rightarrow$ gaussien
- Le prior : $p(\mathbf{w})$ est gaussien (dans le cadre de la ridge régression)
- Le posterior : $p(\mathbf{w}|D) = \frac{p(\mathbf{w})p(D|\mathbf{w})}{p(D)} \Rightarrow$ gaussien

Mais si on se passait de l'estimation de $\mathbf{w}$:

- Ce qu'on veut : $p(y|\mathbf{x}, D)$
- Mais $p(y|\mathbf{x}, D) = \int_{\mathbf{w}} p(y|\mathbf{x}; \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w}$
- Or tous les termes sont gaussiens

$\Rightarrow$ $p(y|\mathbf{x}, D)$ est gaussien !

- Donc $p(y|\mathbf{x}, D) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, il suffit d'estimer $\boldsymbol{\mu}$ et Σ pour prédire y.

En résumé

Soit $D = \{\mathbf{x}^i, y^i\}_{i=1}^N$ nos données et $\mathbf{x}_t^1 \dots \mathbf{x}_t^T$ les points que l'on veut inférer.
On a établit que (en simplifiant en fixant la moyenne à 0) :

$$p\left(\begin{pmatrix} y^1 \\ \vdots \\ y^N \\ y_t^1 \\ \vdots \\ y_t^T \end{pmatrix} | \mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N, \mathbf{x}_t^1, \dots, \mathbf{x}_t^T\right) \sim \mathcal{N}(0, \Sigma),\ \Sigma = \begin{bmatrix} K & K_\star \\ K_\star^t & K_{\star\star} \end{bmatrix}$$

Alors

$$p(y_t^1 \dots y_t^T | \mathbf{y}, \mathbf{x}^1, \dots \mathbf{x}_t^T) \sim \mathcal{N}(K_\star^t K^{-1}\mathbf{y}, K_{\star\star} - K_\star^t K^{-1} K_\star)$$

Mais comment obtenir:

- K les covariances entre les points d'entraînement ?
- $K_\star$ les covariances entre entraînement et test ?
- $K_{\star\star}$ les covariances entre test ?

Matrice de covariance = Kernel !

Ce que l'on veut pour la matrice de covariance :

- qu'elle soit symétrique ! (i influence j comme j influence i)
- deux points "similaires" doivent avoir une corrélation forte : $Cov(\mathbf{x}^i, \mathbf{x}^j)$ grand
- deux points "dissimilaires" doivent avoir une corrélation faible : $Cov(\mathbf{x}^i, \mathbf{x}^j)$ petit
- qu'elle soit semi-défini positive
- $Cov(\mathbf{x}^i, \mathbf{x}^i)$ doit dénoté la variance en ce point

$\Rightarrow$ Très similaire à la notion de noyaux en SVM ! Autant utiliser une fonction noyau pour encoder la covariance ...

Covariances typiques :

- Squared Exponential : $K(\mathbf{x}^1, \mathbf{x}^2) = \sigma^2 e^{-\frac{1}{2}\left(\frac{\|\mathbf{x}^1 - \mathbf{x}^2\|^2}{\lambda}\right)}$
- Linéaire : $K(\mathbf{x}^1, \mathbf{x}^2) = \lambda + \langle \mathbf{x}^1, \mathbf{x}^2 \rangle$
- Periodic : $K(\mathbf{x}^1, \mathbf{x}^2) = \sigma^2 e^{-\frac{2sin^2\left(\frac{\|\mathbf{x}^1 - \mathbf{x}^2\|}{2}\right)}{\lambda^2}}$

Et si on introduit du bruit ?

Bruit additif gaussien

- On observe $y^i = f(\mathbf{x}^i) + \epsilon_i$, avec ϵ_i indépendant et suivant $\mathcal{N}(0, \sigma^2)$
- La covariance est changée en $\hat{\Sigma}$:

$$\hat{\Sigma}_{ij} = \mathbb{E}[(f(\mathbf{x}^i) + \epsilon_i)(f(\mathbf{x}^j) + \epsilon_j)] = \mathbb{E}[f(\mathbf{x}^i)f(\mathbf{x}^j)] + \mathbb{E}[f(\mathbf{x}^j)]\mathbb{E}[\epsilon_i] + \mathbb{E}[f(\mathbf{x}^i)]\mathbb{E}[\epsilon_j] + \mathbb{E}[\epsilon_i\epsilon_j]$$

- Pour $i \neq j$, $\mathbb{E}[\epsilon_i] = 0, \mathbb{E}[\epsilon_i\epsilon_j] = \mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j] = 0$

$$\hat{\Sigma}_{ij} = \mathbb{E}[f(\mathbf{x}^i)f(\mathbf{x}^j)] = \Sigma_{ij}$$

- Pour $i = j$, $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2] = \sigma^2$

$$\hat{\Sigma}_{ii} = \mathbb{E}[f(\mathbf{x}^i)f(\mathbf{x}^i)] + \mathbb{E}[\epsilon_i^2] = \Sigma_{ii} + \sigma^2$$

- Donc $\hat{\Sigma} = \Sigma + \sigma^2 \boldsymbol{I}$

Formule de la régression GP dans le cas général

$$p(y_t^1 \dots y_t^T | \mathbf{y}, \mathbf{x}^1, \dots \mathbf{x}_t^T) \sim \mathcal{N}(K_\star^t (K + \sigma^2 \boldsymbol{I})^{-1} \mathbf{y},\ K_{\star\star} - K_\star^t (K + \sigma^2 \boldsymbol{I})^{-1} K_\star)$$

$\Rightarrow$ Régression à noyaux !

- Mais avec l'information sur l'incertitude liée à la prédiction !

Définition formelle des Processus Gaussien (GP)

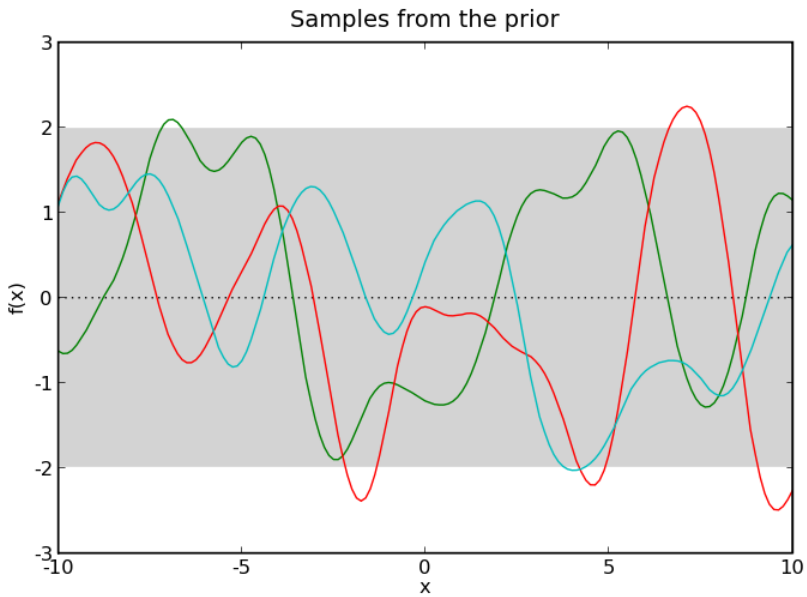
Définition

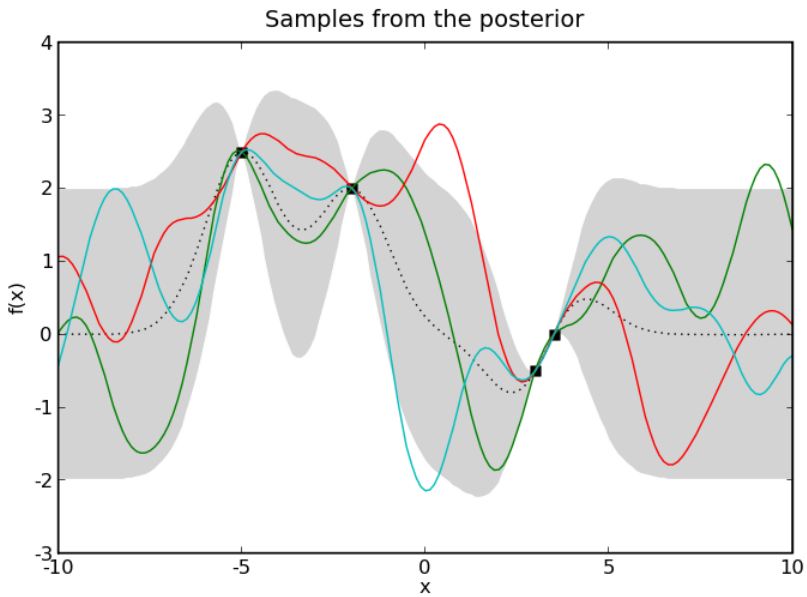
- La fonction f dont on cherche la prédiction est vue comme une collection de points $\boldsymbol{f}$ (les mesures en différents points) potentiellement infinie.
- Un processus gaussien est une collection de variables aléatoires (potentiellement infinie) tels que la distribution jointe de tout sous-ensemble de ces variables est une gaussienne multivariée :

$$f \sim GP(\mu, k)$$

  avec $\mu(\mathbf{x})$ et $k(\mathbf{x}^1, \mathbf{x}^2)$ sont les fonctions de moyenne et de covariance.
- On cherche à estimer la distribution $P(\boldsymbol{f}_t|\mathbf{x}_t, D)$ en utilisant un prior GP : $P(\boldsymbol{f}|\mathbf{x}) \sim \mathcal{N}(\mu, \Sigma)$ et en le conditionnant par les données d'entraînement D afin de modéliser la distribution jointe $\boldsymbol{f}$ des points d'entraînement et $\boldsymbol{f}_t$ les points de test.

Exemples


Samples from the prior
f(x)
x
3
2
1
0
-1
-2
-3
-10
-5
0
5
10


Samples from the posterior
f(x)
x
4
3
2
1
0
-1
-2
-3
-10
-5
0
5
10

Conclusion

Les processus gaussiens

- sont relativement puissants sous certaines conditions
- sont adaptables à beaucoup de tâches (classification, non supervisé, active learning, . . .)
- donnent une mesure d'incertitude liée à la prédiction
- Mais temps de calcul en $O(N^3)$ avec inversion de matrice !
- Les hyper-paramètres sont cachés dans le noyau . . .

Références:

Cours ETHZ
Cours Cornell U., très bonne intro vidéo
Très bon livre de Rasmussen et Williams