Fondements de l'apprentissage machine

Automne 2014

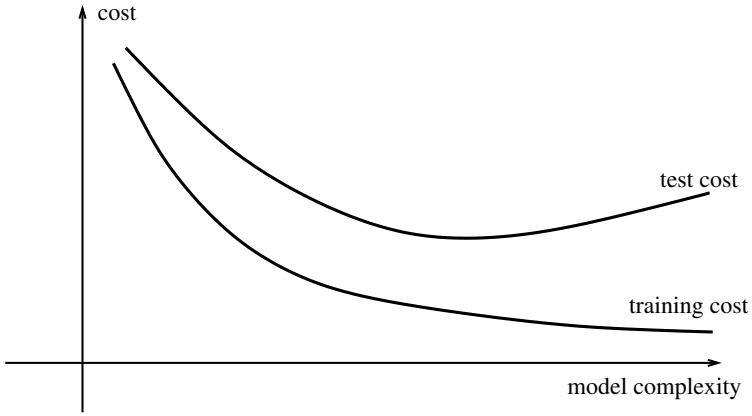
Roland Memisevic

Leçon 5

Plan

- Sur-apprentissage et moyens pour le prévenir
- Raisonnement bayésien
- Distributions a priori conjuguées
- Évidence et sélection du modèle

Sur-apprentissage et réglage de la capacité



Aspects du sur-apprentissage

- Biais/variance
- Taille de l'espace des hypothèses
- Dimension VC
- Malédiction de la dimensionalité

Prévenir le sur-apprentissage

1. weight decay (terme de pénalité)
2. Arrêt prématuré (early stopping)
3. Ajouter des pseudo-comptes (lors de l'estimation d'une distribution discrète)
4. Choisir une classe de modèles restrictifs

- Ceux-ci sont également appelés "lissage" ou "régularisation"
- Les paramètres qui contrôlent la complexité du modèle (par exemple le nombre de pseudo-comptes ou le multiplieur λ pour weight decay) sont appelés hyper-paramètres.

Réglage des hyper-paramètres

- Il existe différentes approches analytiques pour choisir de bonnes valeurs pour les hyper-paramètres : BIC, AIC, MDL, dimension VC.
- L'approche la plus commune et pratiquement éprouvée est de mettre de côté des données de validation pour évaluer les choix des valeurs des hyper-paramètres après l'apprentissage.
- Une variante plus commune, qui nous permet d'utiliser toutes les données d'entraînement, est la K-validation croisée : Partitionner les données d'entraînement en K groupes. Utilisez $K-1$ groupes pour l'entraînement et l'autre pour la validation à chaque fois.

Biais inductif

- Des connaissances sur la tâche à accomplir peuvent également être utilisées pour la régularisation.
- Ces connaissances peuvent prendre différentes formes. Par exemple, nous pouvons savoir que
  - les dépendances entre les entrées et les sorties sont linéaires ; quadratiques ; sinusoïdales, etc.
  - les classes sont séparées par de grandes marges
  - les données sont structurées comme une séquence ; un arbre ; une grille, etc.
- Il n'y a pas d'apprentissage sans faire des suppositions ("no free lunch").

Modélisation bayésienne

- La modélisation bayésienne est une façon d'apprendre différente de toutes les méthodes dont nous avons discutées jusqu'ici.
- Elle *n'est pas* basée sur l'optimisation d'une fonction de perte.
- Elle est basée sur l'inférence probabiliste, traitant les données et les paramètres comme des variables aléatoires.
- Dans le cadre de la modélisation bayésienne, l'obligation de faire des suppositions pour être en mesure d'apprendre se reflète dans l'exigence de définir une distribution *a priori* sur les paramètres.

Modélisation bayésienne

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}$$

- Pour obtenir la distribution a posteriori sur les paramètres $\boldsymbol{\theta}$, il faut multiplier la vraisemblance, $p(D|\boldsymbol{\theta})$, avec la distribution a priori, $p(\boldsymbol{\theta})$, et normaliser.
- (Cela nécessite d'interpréter une probabilité comme quelque chose d'autre qu'une fréquence relative, ce qui a causé de nombreux débats philosophiques, les bayésiens acceptant cela, au contraire des fréquentistes.)

Modélisation bayésienne

- Il faut fournir une distribution a priori qui reflète la croyance du modélisateur avant d'avoir vu les données.
- Au lieu d'une seule ≪bonne≫ réponse, la modélisation bayésienne donne une distribution a posteriori, ce qui reflète la croyance du modélisateur après avoir vu les données.
- Dans un contexte où nous obtenons des données de façon séquentielle, nous pouvons traiter $p(\boldsymbol{\theta}|\mathcal{D})$ comme distribution a priori et la mettre à jour lorsque plus de données arrivent.
- Dans ce cas, on écrit parfois $p(\boldsymbol{\theta}|\mathcal{D})$ à la fois pour la distribution a posteriori et la distribution a priori, où $\mathcal{D}$ est un ensemble vide pour cette dernière.

Distribution prédictive

- Pour appliquer le modèle à des données de test (dans un problème supervisé), il faut utiliser les règles de probabilité pour calculer la probabilité sur des sorties sachant l'entrée et les données d'entraînement $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{t}_1), \ldots, (\mathbf{x}_N, \mathbf{t}_N)\}$
- La distribution prédictive s'avère être une moyenne de modèles pondérés par leur postérieur :

$$p(\mathbf{t}|\mathcal{D}, \mathbf{x}) = \int p(\mathbf{t}|\boldsymbol{\theta}, \mathbf{x})p(\boldsymbol{\theta}|\mathcal{D})\,\mathrm{d}\boldsymbol{\theta}$$

Distribution prédictive

- Cela donne souvent des prédictions plus précises que celles d'un modèle dont les paramètres ont été estimés par une estimation ponctuelle.
- La philosophie de la modélisation bayésienne peut être résumée comme ceci : "Mettez tout ce que vous savez sur la table, puis utilisez les règles de probabilité pour répondre à n'importe quelle question".
- En d'autres termes, nous spécifions une distribution jointe sur toutes les quantités d'intérêt. Le reste est inférence probabiliste.
- Le désavantage : la modélisation bayésienne est généralement exigeante, à la fois mathématiquement et en ce qui concerne les calculs. De plus, elle nécessite souvent une inférence approximative ou des méthodes d'échantillonnage.

Distributions conjuguées

- Les dérivations peuvent se simplifier si le produit :

$$p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

  a la même forme que les composants $P(\mathcal{D}|\boldsymbol{\theta})$ et $p(\boldsymbol{\theta})$
- Une distribution a priori qui satisfait cette propriété, pour une fonction de vraisemblance donnée, est appelée distribution a priori conjuguée.
- Des exemples de distributions conjuguées comprennent :
  - la distribution bêta (conjuguée à la distribution Bernoulli),
  - la distribution Dirichlet (pour la distribution multinomiale),
  - la distribution Wishart (pour la variance de la gaussienne).

Bernoulli et distribution bêta

- Rappelons que la **distribution Bernoulli** peut être écrite

$$p(x|\mu) = \mu^x(1-\mu)^{1-x}$$

  où x est 0 ou 1.
- La vraisemblance de l'ensemble $\mathcal{D}$, composé de m entrées 1's et $l = N - m$ entrées 0, est

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} = \mu^{\sum_n x_n}(1-\mu)^{\sum_n (1-x_n)} = \mu^m(1-\mu)^l$$

- (Notez que si μ est une variable aléatoire, nous écrivons maintenant $p(x|\mu)$ au lieu de $p(x;\mu)$)

Bernoulli et distribution bêta

- La distribution conjuguée de la Bernoulli est

La Distribution Bêta

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

La moyenne et la variance sont

$$\mathbb{E}[\mu] = \frac{a}{a+b}, \quad \text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

où a and b sont des paramètres de valeur réelle positive, et $\Gamma(\cdot)$ est la fonction Gamma : $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}\text{d}u$

- Preuve :

$$p(\mu|\mathcal{D},a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

Distribution prédictive et pseudo-comptes

- $p(\mu|\mathcal{D},a,b)$ est elle-même une distribution bêta, la constante de normalisation nous étant donnée par :

$$\frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}$$

- La **distribution prédictive** est

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})\,\text{d}\mu = \int \mu p(\mu|\mathcal{D})\,\text{d}\mu$$

  (donc elle est l'espérance de la distribution a posteriori)

Distribution prédictive et pseudo-comptes

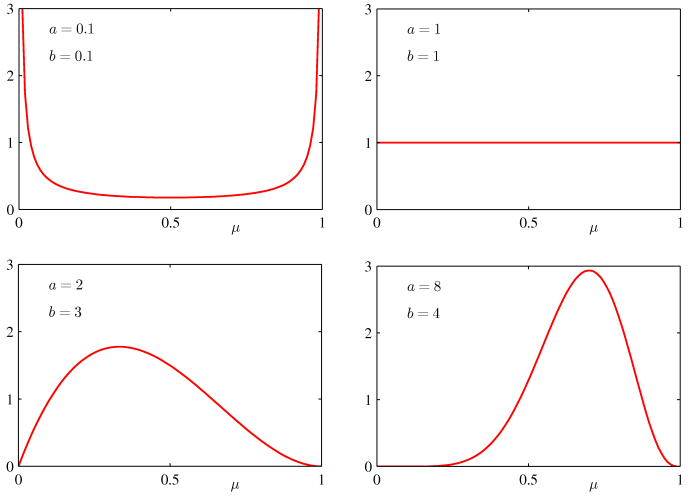
- L'entraînement par un ensemble de m 1 et de l 0 nous donne :

$$p(x = 1|\mathcal{D}) = \frac{m + a}{m + a + l + b}$$

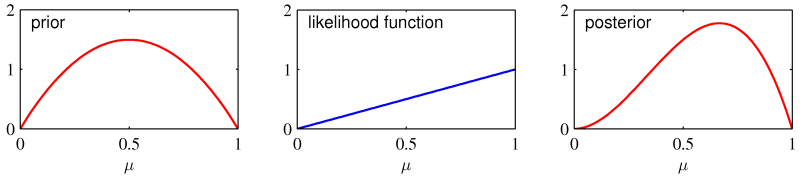
- Cela signifie qu'utiliser la distribution a priori $\mathrm{Beta}(\mu|a, b)$ conduit à l'ajout de **pseudo-comptes** a et b à l'estimation du maximum de vraisemblance, qui est

$$\mu = \frac{m}{m + l}$$

Exemples de distributions bêta



Distribution a posteriori $\propto$ vraisemblance $\times$ distribution a priori



- Notez que la distribution a priori et la vraisemblance prennent la même forme fonctionnelle (comme fonction de μ)
- Dans cet exemple, la distribution a priori est $\mathrm{Beta}(2, 2)$, avec $m = N = 1$, $l = N - m = 0$, donc la distribution a posteriori est $\mathrm{Beta}(3, 2)$

Distributions multinoulli et Dirichlet

- Rappelons que la **distribution discrète** peut être écrite

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

où $\mathbf{x}$ est en codage orthogonal.

- (Ceci est bien entendu une généralisation de la distribution Bernoulli)
- La vraisemblance des données $\mathcal{D}$, representées par la matrice $\mathbf{X}$, est

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\sum_n x_{nk}} =: \prod_{k=1}^{K} \mu_k^{m_k}$$

Distributions multinoulli et Dirichlet

- La distribution conjuguée de la distribution multinoulli est

La Distribution Dirichlet

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \quad \alpha_0 = \sum_k \alpha_k$$

La moyenne et la variance sont

$$\mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0}, \quad \mathrm{var}[\mu_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

où les α_k sont des paramètres de valeur réelle positive.

- Preuve :

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

Distribution multinoulli et Dirichlet

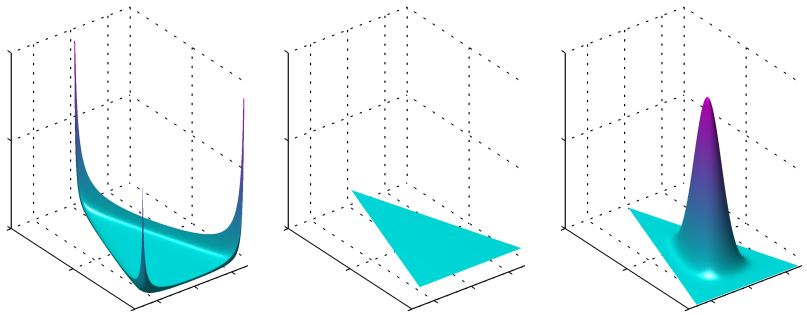
- Comme $p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha})$ est elle-même une distribution Dirichlet, la constante de normalisation nous est donnée par :

$$\frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1), \ldots, \Gamma(\alpha_K + m_K)}$$

- Le moyenne de la distribution a posteriori (ainsi que les paramètres de la distribution prédictive) sont également donnés par des **pseudo-comptes** ajoutés à l'estimation du maximum de vraisemblance :

$$\mathbb{E}[\mu_k] = \frac{m_k + \alpha_k}{\alpha_0}$$

Des exemples de la distribution Dirichlet



- Distributions Dirichlet avec tous les $\alpha_1 = \ldots = \alpha_K = \{0.1(\text{gauche}), 1(\text{milleu}), 10(\text{droite})\}$
- La distribution Dirichlet est confinée au *simplexe*.

Distributions conjuguées pour la gaussienne 1-D

- La distribution conjuguée pour la moyenne μ d'une distribution gaussienne 1-D dont la *variance est constante* est la gaussienne $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ avec des paramètres μ_0, σ_0^2
- Pour la distribution postérieure, la moyenne est $\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\frac{\sum_n x_n}{N}$ et la précision (inverse de la variance) est $\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$
- La distribution conjuguée pour la précision λ d'une gaussienne 1-D dont la *moyenne est constante* est la distribution Gamma : $\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$
- La distribution conjuguée pour la moyenne et la précision d'une gaussienne 1-D est le produit $p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-}1)\mathrm{Gam}(\lambda|a, b)$

Distributions conjuguées pour la gaussienne multivariée

- La distribution conjuguée pour la moyenne $\boldsymbol{\mu}$ d'une distribution gaussienne multivariée dont la *matrice de covariance est constante* est une gaussienne multivariée $p(\boldsymbol{\mu})$
- La distribution conjuguée pour la matrice de précision Λ (inverse de la matrice de covariance) d'une gaussienne multivariée dont la *moyenne est constante* est la distribution Wishart.
- La distribution conjuguée pour la moyenne et la matrice de précision d'une gaussienne multivariée est le produit d'une gaussienne avec une distribution Wishart.

Régression linéaire bayésienne

- Un choix commode est

$$p(\mathbf{w}) = \mathcal{N}\big(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\big)$$

pour un certain paramètre de précision α.

- La distribution postérieure pour des données $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$, representées par les matrices $\mathbf{X}$, $\mathbf{t}$, est

$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}\big(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N\big)$$

où

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{X}^{\mathrm{T}} \mathbf{t}$$

et

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

.

Régression linéaire bayésienne

- La régression linéaire peut être définie par la gaussienne conditionnelle

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}\big(t|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}\big)$$

Cela est une fonction exponentielle d'une fonction quadratique de $\mathbf{w}$.

- Pour cette raison, la distribution conjuguée est également une gaussienne

$$p(\mathbf{w}) = \mathcal{N}\big(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\big)$$

- Pour le voir, complétez le carré dans l'exponentielle, ou utilisez les identités gaussiennes (par exemple, Bishop page 93).

Distribution prédictive

- La distribution prédictive est

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta)\, \mathrm{d}\mathbf{w}$$

- Ceci peut être simplifié à

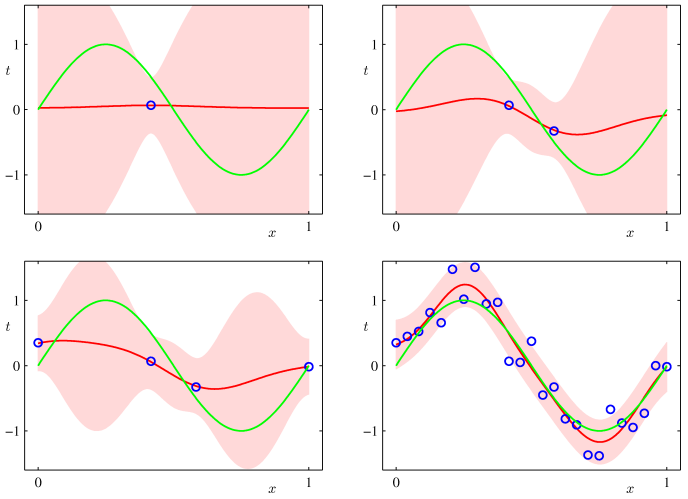
$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\big(t|\mathbf{m}_N^{\mathrm{T}}\mathbf{x}, \sigma_N^2(\mathbf{x})\big)$$

où

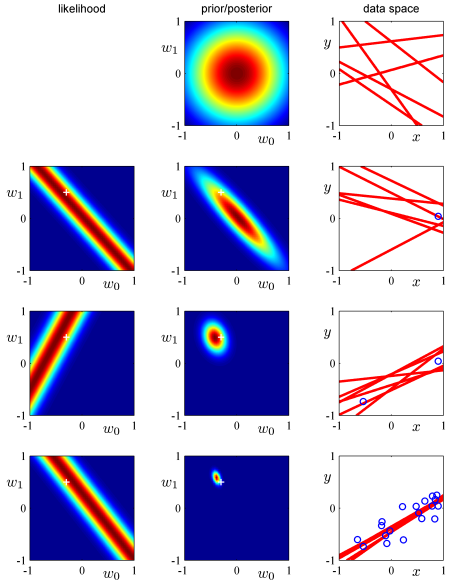
$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \mathbf{x}^{\mathrm{T}}\mathbf{S}_N\mathbf{x}$$

- La variance dépend de $\mathbf{x}$.
- La moyenne a la même forme que pour le modèle ridge regression.

Distribution prédictive : exemple



Régression bayésienne 1-D : exemple revisité



Estimation MAP et régularisation

- Le maximum de la distribution a posteriori s'appelle estimation maximum a-posteriori (MAP).
- Pour n'importe quel modèle de régression avec une distribution a priori gaussienne de moyenne zéro et dont la matrice de covariance est diagonale, l'estimation MAP est identique à la solution de ridge regression :

$$\begin{aligned}\log p(\mathbf{w}|\mathcal{D}) &= \text{const} + \log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}) \\ &= \text{const} - \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - f(\mathbf{x}_n;\mathbf{w})\right)^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\end{aligned}$$

Évidence et sélection de modèle

- La constante de normalisation

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

  est également connue comme évidence.
- En marginalisant sur les paramètres, elle représente *la probabilité de l'ensemble de données étant donné le type de modèle*.
- Son application principale est la sélection de modèle :

Évidence et sélection de modèle

- L'évidence nous permet de choisir un modèle parmi un ensemble $\mathcal{M}_1, \ldots, \mathcal{M}_L$ de modèles (par exemple, régression polynomiale d'ordres différents).
- Nous définissons une distribution a priori $p(\mathcal{M}_i)$ sur les classes and calculons les distributions a posteriori

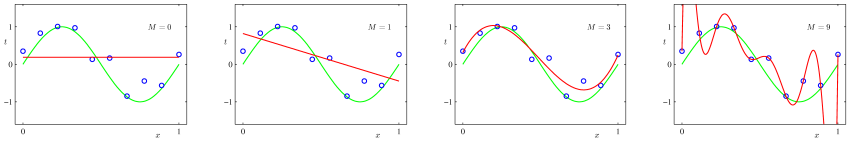
$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)$$

où $p(\mathcal{D}|\mathcal{M}_i)$ est l'évidence pour le type du modèle $\mathcal{M}_i$. Alternativement, nous pouvons faire des prédictions en utilisant les règles de probabilité et en marginalisant sur les modèles.

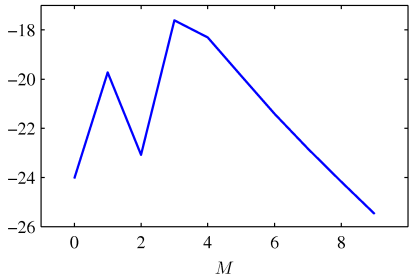
- Pour des gaussiennes et la régression linéaire, on peut calculer $p(\mathcal{D})$ en forme fermée (Bishop, page 169).

Régression polynomiale revisitée

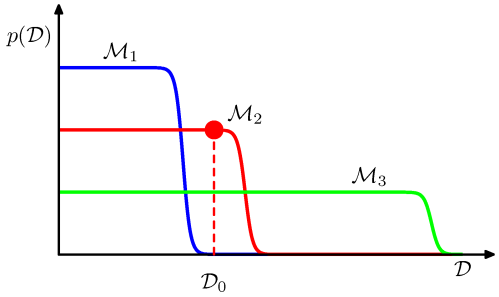
L'exemple de la régression polynomiale de la leçon 2 :



La log-évidence pour différents ordres, M :



Évidence : intuition

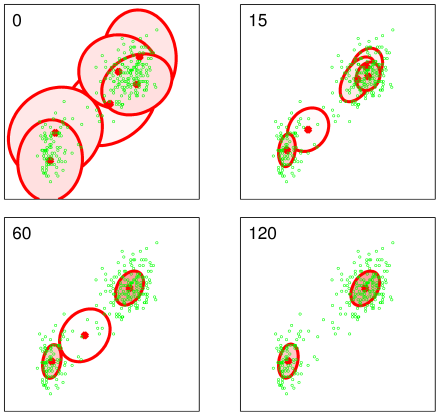


- Des modèles simples attribuent beaucoup de masse de probabilité à un petit nombre d'ensembles de données.
- Des modèles complexes diffusent leur masse de probabilité sur de nombreux ensembles de données.

Modèles non-linéaires

- L'inférence bayésienne est possible dans une forme fermée pour certains modèles simples seulement.
- Pour des modèles plus compliqués, il faut utiliser l'inférence approximative et l'échantillonnage.
- Quelques exemples dans le livre Bishop :
- Régression logistique (approximation Laplace), page 217
- Régression logistique (inférence variationnelle), page 498
- Réseaux de neurones (approximation Laplace), page 277
- Mélange de gaussiennes (inférence variationnelle), page 474

Un modèle de mélange bayésien peut déduire le nombre de grappes à partir des données



- Itération indiquée en haut à gauche.