Fondements de l'apprentissage machine
II. Minimisation du Risque et Théorie de la Décision

Pierre-Luc Bacon / Vincent Taboga

Références

Dans PML:

- Section 2.5
- Section 4.2
- Section 4.5
- Section 5.4.4
- Section 9.2

Minimisation du risque

Nous voyons l'apprentissage comme un problème d'estimation d'une fonction. Cette approximation est celle qui minimise le risque:

$$\text{minimiser}_{f \in \mathcal{F}} \; R(f)$$

où $\mathcal{F}$ est une famille de fonctions et $R(f) \doteq \mathbb{E}\left[l(Y, f(X))\right]$ Puisque le vrai risque ne peut pas être évalué directement et exactement, nous utilisons plutôt le risque empirique:

$$\hat{R}(f) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} l(y_i, f(x_i)) \quad .$$

Fonction de perte de substitution

Dans le cas de la classification, nous pourrions par exemple choisir la fonction de perte 0-1:

$$l_{01}(y_i, f(x_i)) \doteq \begin{cases} 0 & \text{si } y_i = f(x_i) \\ 1 & \text{si } y_i \neq f(x_i) \end{cases}$$

Toutefois la fonction de perte 0-1 est non lisse, ce qui rend l'optimisation difficile. Nous préférons plutôt utiliser une **fonction de perte de substitution** (surrogate loss function), qui est souvent choisie comme étant une borne supérieure convexe à la fonction d'origine. Cette fonction substitution est généralement plus facile à optimiser.

Exemples

Dans un contexte probabiliste de classification, nous pourrions plutôt choisir une fonction sigmoide de sorte que:

$$p(y|x) = \frac{1}{1 + e^{-yf(x)}} \quad , \quad \text{et} \quad l_{\text{ll}}(y, f(x)) \doteq -\log p(y|x) = \log(1 + e^{-yf(x)})$$

où $f(x)$ représente la quantité appelée **logit** (ou logg-odds). La minimisation de la **log de vraisemblance négative** est donc équivalente à la minimisation d'une borne supérieure pour la fonction de perte 0-1. Un autre choix possible de fonction de substitution est la **fonction de perte à charnière** (hinge loss):

$$l_{\text{hinge}}(y, f(x)) \doteq \max(0, 1 - yf(x)) \quad ,$$

qui est différentiable seulement **par pièces** (mais demeure une borne supérieure convexe).

Examples de bornes supérieures convexes

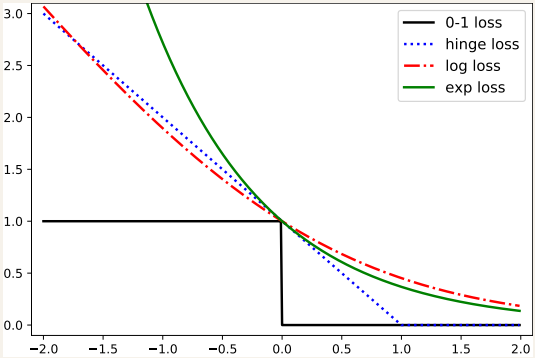


Figure: Deux exemples de fonctions de pertes de substitution agissant comme bornes supérieures convexes à la fonction de perte 0-1.

Estimateur du maximum de la vraisemblance

Il se trouve que le principe de minimisation du risque empirique coïncide, pour un choix de fonction de perte approprié, avec l'estimateur du maximum de la vraisemblance. En supposant que nos données soient indépendamment et identiquement distribuées, le logarithme d'un produit se transforme alors en une somme de logarithmes telle que:

$$\hat{R}(f) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|f(x_i)) \ .$$

La forme exacte de $p(y_i|f(x_i))$ dépend de nos suppositions sur la distribution des données.

Régression

Dans le contexte de régression, nous utilisons principalement **l'erreur de perte quadratique**: $l_2(y, \hat{y}) \doteq \|y - \hat{y}\|_2^2$. Le risque empirique de cette fonction de perte coïncide alors avec **l'erreur quadratique moyenne (EQM)** (mean squared error (MSE)):

$$MSE(\theta) = \hat{R}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \|y_i - f(x_i; \theta)\|_2^2 \ .$$

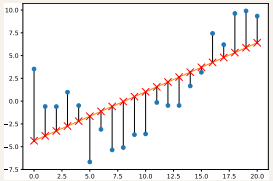


Figure: En régression par la méthode des moindres carrés, nous voulons une fonction qui minimise la somme des **résidus**

Connection entre L2 et EQM

Nous pourrions alors choisir de modéliser la distribution conditionnelle des cibles étant donné les entrées par une distribution gaussienne (ou normale). Plus précisément, choisissons de fixer la variance à une valeur σ^2 et définissons la moyenne de cette distribution comme étant la sortie d'une fonction f. Nous avons alors:

$$p(y|x;\theta) = \mathcal{N}(y|f(x;\theta),\sigma^2) \ , \ \text{avec} \ \mathcal{N}(y|\mu,\sigma^2) \doteq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \ .$$

La log vraisemblance négative est alors:

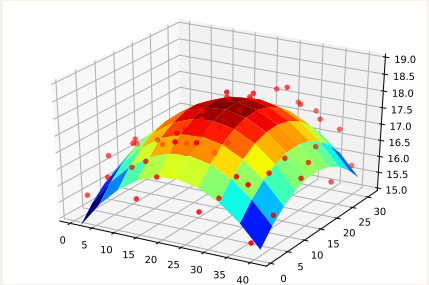
$$\begin{aligned} NLL(\theta) = \hat{R}(\theta) &= -\frac{1}{N}\sum_{i=1}^{N} \log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - f(x_i;\theta)\right)^2\right)\right) \\ &= \frac{1}{2\sigma^2} MSE(\theta) + \text{constante} \ . \end{aligned}$$

Modèles de régression

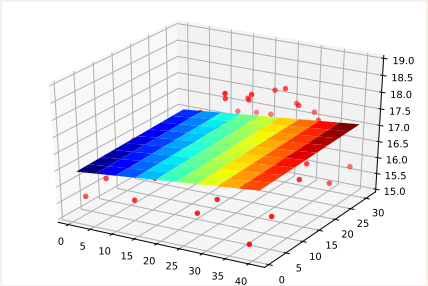
En régression linéaire (régression linéaire multiple), nous modélisons la cible par une combinaison linéaire des traits d'un exemple: $f(x; w, b) = w^\top x + b$ avec $x \in \mathbb{R}^d, w \in \mathbb{R}^d$ et $b \in \mathbb{R}$. Les **parametètres** de ce modèle sont w et b. Nous utilisons parfois le symbole θ pour désigner l'ensemble de tous les paramètres propre à un modèle.

En **régression polynomiale**, nous appliquons d'abord une transformation non linéaire aux traits d'un exemple par l'entremise d'une fonction $\phi : \mathbb{R} \to \mathbb{R}^{k+1}$ où k est le degré du polynôme: $\phi(x) = [1, x, x^2, \ldots, x^k]$. La fonction ϕ projette (maps) les données dans un **espace de redescription** (feature space).

Espace de redescription: intuition



(a) Régression polynomiale: régression linéaire dans un espace de redescription. $f(x; w, b) = b + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$



(b) Régression linéaire sans fonction de mappage: $f(x; b, w) = w_0 + w_1 x_1 + w_2 x_2$

Figure: Modélisation de la température en fonction de la position en 2D dans une chambre.

Capacité et degré du polynôme

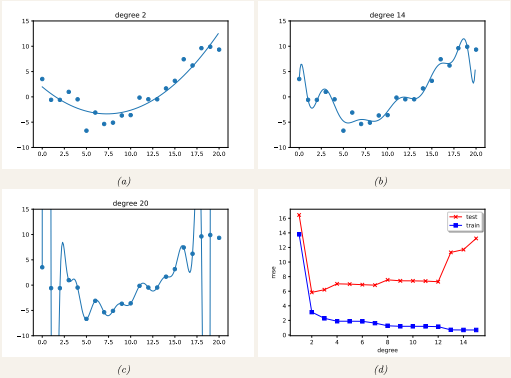


Figure: Lorsque $k = N - 1$ nous avons un paramètre par exemple et pouvons **interpoler** les données parfaitement. Bien que l'EQM soit 0, il ne s'agit pas nécessairement d'un modèle désirable du point de vue de la généralisation.

Généralisation: vocabulaire

Rappelons-nous, le risque est défini comme étant: $R(\theta) = \mathbb{E}\left[l(Y, f(Y;\theta)\right]$, et le risque empirique, pour un ensemble de données $\mathcal{D}$ est: $\hat{R}(\theta;\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} l(y_i, f(x_i;\theta))$. La différence $R(\theta) - \hat{R}(\theta;\mathcal{D}_{\text{train}})$ est appelée **l'écart de généralisation** (generalization gap).

Le risque d'un modèle peut être estimé par l'entremise du risque empirique sur un **ensemble de test**: $\hat{R}(\theta;\mathcal{D}_{\text{test}})$. L'erreur de test est caractérisée par une **courbe en U** (u-shaped curve).

Un modèle peut souffrir de **surapprentissage** (overfitting, $k \gg 1$ dans l'exemple précédent) ou de **sous-apprentissage** (underfitting, $D = 1$). La complexité du modèle est appropriée dans l'exemple précédent lorsque $k = 2$.

Un troisième **ensemble de validation** est utilisé pour la **sélection de modèles** (model selection), alors que $\mathcal{D}_{\text{test}}$ "doit être gardé dans un coffre-fort jusqu'à la fin".

No free lunch theorem

Il n'existe pas de modèle parfait qui fonctionne optimalement pour tous les problèmes (énoncé plus formel dans Wolpert en 1996): le **no free lunch theorem**.

Le phénomène de l'apprentissage dépend des **biais inductifs** (inductive biases) exprimés dans nos modèles – les suppositions implicites ou explicites de notre modèle. Ces biais inductifs sont spécifiques à une classe de problèmes.

MAP et MLE

D'un point de vue bayésien, l'estimateur du maximum de la vraisemblance est:

$$\hat{\theta}_{\text{MLE}} \doteq \arg\max_{\theta} p(\mathcal{D}|\theta) \ ,$$

où $\mathcal{D}$ est une variable aléatoire. Si les données sont échantillonnées i.i.d:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y_i|x_i, \theta) \ .$$

En prenant le logarithme, nous obtenons le risque empirique:

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(y_i|x_i, \theta) \ .$$

MAP et MLE

L'estimateur du maximum de la vraisemblance est alors:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i=1}^{N} \log p(y_i|x_i, \theta) \ .$$

En adoptant la perspective d'optimisation sur l'apprentissage, nous voulons alors **minimiser** une fonction de mesure de performance. La maximisation de la vraisemblance devient alors un problème de minimisation du logarithme négatif de la vraisemblance:

$$NLL(\theta) \doteq \arg\min_{\theta} -\sum_{i=1}^{N} \log p(y_i|x_i, \theta) \ .$$

MAP et MLE

À l'aide du théorème de Bayes, nous avons:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \ .$$

L'estimateur du maximum a posteriori (MAP) est alors:

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta} \log p(\theta|\mathcal{D}) = \arg\max_{\theta} \log p(\mathcal{D}|\theta) + \log p(\theta) \ .$$

Si nous choisisons une distribution apriori uniforme $p(\theta) \propto 1$, **l'estimateur MAP est alors égal à l'estimateur MLE**.

Théorie de l'information et MLE

prenons le cas non conditionnel du MLE, tel que:

$$\hat{\theta}_{\text{mle}} = \arg\min_{\theta} -\sum_{i=1}^{N} \log p(y_i|\theta) \ .$$

Nous pouvons justifier le principe MLE en montrant que la distribution prédictive $p(y|\hat{\theta}_{\text{MLE}})$ est celle qui s'approche le plus de la distribution empirique des données:

$$p_{\mathcal{D}}(y) \doteq \frac{1}{N} \sum_{i=1}^{N} \delta(y - y_i) \ .$$

Cette distribution empirique est une série de fonctions delta positionnée sur chaque exemple. Nous voulons un modèle $q(y) = p(y|\theta)$ qui est similaire à $p_{\mathcal{D}}$.

Théorie de l'information et MLE

Nous pouvons mesurer la similarité entre deux distributions via la divergence de Kullback Leibler (KL, pas une distance):

$$\begin{aligned} D_{\mathsf{KL}}(p||q) &= \sum_y p(y) \log \frac{p(y)}{q(y)} \\ &= \sum_y p(y) \log p(y) - \sum_y p(y) \log q(y) \ . \end{aligned}$$

Le premier terme est **l'entropie** négative de p que nous écrivons $\mathbb{H}(p)$ et le second est l'entropie croisée entre p et q. $D_{\mathsf{KL}}(p||q) \geq 0$ avec égalité ssi $p = q$.

Théorie de l'information et MLE

Avec $q(y) = p(y|\theta)$ et $p(y) = p_{\mathcal{D}}(y)$, la divergence de KL est:

$$\begin{aligned} D_{KL}(p||q) &= \sum_y (p_{\mathcal{D}}(y) \log p_{\mathcal{D}}(y) - p_{\mathcal{D}}(y) \log q(y)) \\ &= -\mathbb{H}(p_{\mathcal{D}}) - \frac{1}{N} \sum_{i=1}^{N} \log p(y_i|\theta) \\ &= \text{constante} + NLL(\theta) \ . \end{aligned}$$

Cas conditionnel

Le même argument s'applique également au cas conditionnel. Il suffit cette fois d'utiliser la règle en chaîne pour écrire la distribution empirique comme étant:

$$p_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{y}) = p_{\mathcal{D}}(\boldsymbol{y} \mid \boldsymbol{x}) p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta\left(\boldsymbol{x} - \boldsymbol{x}_n\right) \delta\left(\boldsymbol{y} - \boldsymbol{y}_n\right)$$

L'espérance de la divergence de KL est alors:

$$\begin{aligned}
\mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})}\left[D_{\mathbb{KL}}\left(p_{\mathcal{D}}(Y \mid \boldsymbol{x}) \| q(Y \mid \boldsymbol{x})\right)\right] &= \sum_{\boldsymbol{x}} p_{\mathcal{D}}(\boldsymbol{x}) \left[\sum_{\boldsymbol{y}} p_{\mathcal{D}}(\boldsymbol{y} \mid \boldsymbol{x}) \log \frac{p_{\mathcal{D}}(\boldsymbol{y} \mid \boldsymbol{x})}{q(\boldsymbol{y} \mid \boldsymbol{x})}\right] \\
&= \text{const} - \sum_{\boldsymbol{x}, \boldsymbol{y}} p_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{y}) \log q(\boldsymbol{y} \mid \boldsymbol{x}) \\
&= \text{const} - \frac{1}{N} \sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}\right)
\end{aligned}$$

EMV pour la loi de Bernoulli

Soit $\theta = p(Y = 1)$ la probabilité de l'événement "face" dans un tirage au sort. Le EMV pour θ est alors:

$$\begin{aligned}\mathrm{NLL}(\theta) = -\log \prod_{n=1}^{N} p\left(y_n \mid \theta\right) &= -\log \prod_{n=1}^{N} \theta^{\mathbb{I}(y_n=1)}(1-\theta)^{\mathbb{I}(y_n=0)} \\ &= -\sum_{n=1}^{N} \left[\mathbb{I}\left(y_n = 1\right)\log\theta + \mathbb{I}\left(y_n = 0\right)\log(1-\theta)\right] \\ &= -\left[N_1 \log\theta + N_0 \log(1-\theta)\right]\end{aligned}$$

où $N_1 = \sum_{n=1}^{N_{\mathcal{D}}} \mathbb{I}\left(y_n = 1\right)$ et $N_0 = \sum_{n=1}^{N_{\mathcal{D}}} \mathbb{I}\left(y_n = 0\right)$, représentant le nombre de pile et face.

EMV pour la loi de Bernoulli

Puisque nous définissons l'EMV comme étant un minimiseur du logarithme négatif de la vraisemblance, nous posons:

$$D\,\mathrm{NLL}(\theta) = \frac{-N_1}{\theta} + \frac{N_0}{1-\theta} = 0 \ ,$$

où $D\,\mathrm{NLL}(\theta)$ signifie la dérivée de la fonction *NLL* évaluée à la valeur θ. La solution à ce système d'équations est alors:

$$\hat{\theta}_{\mathrm{mle}} = \frac{N_1}{N_0 + N_1}$$

.

Distribution catégorique

Une généralisation de l'exemple précédent est celle d'une distribution catégorique (un cas spécial de la loi multinomiale) qui nous sert à modéliser le problème de classification à plusieurs classes. D'un point de vue probabiliste, nous pouvons imaginer avoir un dé à C côtés que nous lançons N fois. Le n-ième événement est représenté par la variable aléatoire $Y_n \in \{1, \ldots, C\}$ où $Y_n \sim \mathrm{Cat}(\boldsymbol{\theta})$ et:

$$\mathrm{Cat}(\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)} \ ,$$

ce qui signifie que $p(y = c|\boldsymbol{\theta}) = \theta_c$. De plus, les paramètres sont contraints à respecter $0 \leq \theta_c \leq 1$ et $\sum_{c=1}^{C} \theta_c = 1$. Dans le cadre de la classification, nous avons un paramètre θ_c par classe.

EMV pour la distribution catégorique

La présence de contraintes sur les paramètres de la distribution catégorique demande l'utilisation de la méthode des multiplicateurs de Lagrange. Nous faisons face au problème:

$$\begin{aligned}
&\text{minimiser} \quad \mathrm{NLL}(\boldsymbol{\theta}) \triangleq -\sum_k N_k \log \theta_k \\
&\text{tel que} \quad \mathbf{0} \leq \boldsymbol{\theta} \leq \mathbf{1} \\
&\text{ainsi que} \quad \mathbf{1}^\top \boldsymbol{\theta} = 1 \ .
\end{aligned}$$

où N_k est le nombre de fois que l'événement $Y = k$ est observé dans les données $\mathcal{D} = \{y_n : n = 1 : N\}$. Le lagrangien correspondant à ce problème est:

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) \triangleq -\sum_k N_k \log \theta_k - \lambda \left(1 - \sum_k \theta_k\right)$$

EMV pour la distribution catégorique

Les dérivées partielles sont:

$$D_\lambda L(\boldsymbol{\theta}, \lambda) = 1 - \sum_k \theta_k = 0$$

$$D_{\theta_k} L(\boldsymbol{\theta}, \lambda) = -\frac{N_k}{\theta_k} + \lambda = 0$$

Nous trouvons alors que $\theta_k = \frac{N_k}{\lambda}$ et éliminons λ par l'entremise la contrainte $\mathbf{1}^\top \boldsymbol{\theta} = 1$. Plus précisément, puisque $N_k = \lambda \theta_k$, considérons alors $\sum_k N_k = \lambda \sum_k \theta_k$ et donc $\sum_k N_k = \lambda$. En résumé, nous avons:

$$\hat{\theta}_k = \frac{N_k}{\lambda} = \frac{N_k}{\sum_k N_k} = \frac{N_k}{|\mathcal{D}|} \ .$$

Interpretation: l'EMV est simplement la fréquence de chaque type d'événement

EMV pour la régression linéaire

Le modèle de régression linéaire considéré précédement était:

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}\left(y \mid f_\mu(\boldsymbol{x}; \boldsymbol{\theta}), f_\sigma(\boldsymbol{x}; \boldsymbol{\theta})^2\right)$$

où $f_\mu(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathbb{R}$ est une fonction retournant la moyenne et $f_\sigma(\boldsymbol{x}; \boldsymbol{\theta})^2 \in \mathbb{R}_+$ en est une qui calcule la variance. Dans le cas où la variance est fixe, nous parlons alors de régression homoscédastique. La **régression linéaire homoscédastique** est alors une méthode qui tente de modéliser la distribution conditionelle sur les cibles tel que:

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}\left(y \mid \boldsymbol{w}^\top \boldsymbol{x} + b, \sigma^2\right) \ ,$$

avec $\boldsymbol{\theta} = \left(\boldsymbol{w}, b, \sigma^2\right)$.

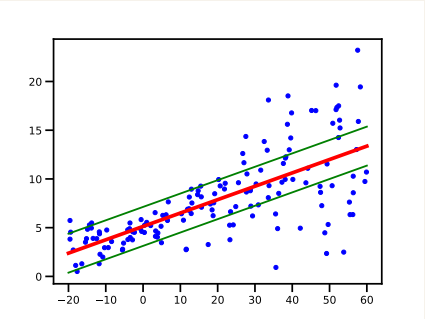
EMV pour la régression linéaire

Si la variance dépend aussi de l'entrée est des paramètres, nous avons un modèle de **régression linéaire hétéroscédastique** sous la forme:

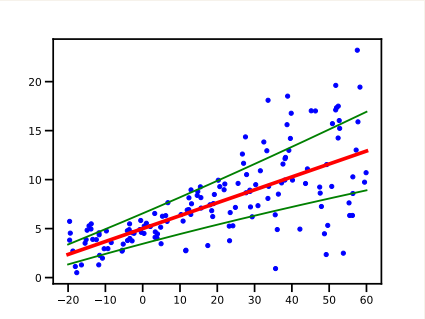
$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}\left(y \mid \boldsymbol{w}_\mu^\top \boldsymbol{x} + b, \sigma_+\left(\boldsymbol{w}_\sigma^\top \boldsymbol{x}\right)\right)$$

avec $\boldsymbol{\theta} = (\boldsymbol{w}_\mu, \boldsymbol{w}_\sigma)$ et $\sigma_+(a) = \log\left(1 + e^a\right)$ pour garantir que la variance prédite soit non négative. Dans un contexte d'apprentissage profond, $\sigma_+ : \mathbb{R} \to \mathbb{R}_+$ est appelée la fonction **softplus**.

Homoscédastique vs hétéroscédastique



(a) Régression linéaire homoscédastique



(b) Régression linéaire hétéroscédastique

Figure: Intervalle de confiance 95% $[\mu(x) - 2\sigma(x), \mu(x) + 2\sigma(x)]$. Le modèle hétéroscédastique exprime l'incertitude dans la prédiction y étant donné x, et non l'incertitude sur les paramètres θ eux-même.

EMV pour la régression linéaire homoscédastique

Choisissons $\boldsymbol{\theta} = \left(\boldsymbol{w}, \sigma^2\right)$ avec une variance fixe (cas homoscédastique):

$$\mathrm{NLL}(\boldsymbol{w}) = -\sum_{n=1}^{N_{\mathcal{D}}} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n \right)^2 \right) \right]$$

Si nous ignorons les quantités ne contenant pas les paramètres d'intérêt, nous obtenons la **somme des carrés des résidus** (SCR/RSS, residual sum of squares) :

$$\mathrm{RSS}(\boldsymbol{w}) \triangleq \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n \right)^2 = \sum_{n=1}^{N} r_n^2$$

avec r_n étant le n-ième résidu.

EMV pour la régression linéaire homoscédastique

Le SCR (RSS) est lié à la l'erreur quadratique moyenne (EQM/MSE) de la manière suivante:

$$\mathrm{MSE}(\boldsymbol{w}) = \frac{1}{N}\,\mathrm{RSS}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right)^2$$

et finalement l'EQM (MSE) est un proche cousin de la racine de l'erreur quadratique moyenne (REQM) puisque:

$$\mathrm{RMSE}(\boldsymbol{w}) = \sqrt{\mathrm{MSE}(\boldsymbol{w})} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right)^2}$$

Cet exercice nous montre donc que le EMV/MLE pour la régression linéaire homoscédastique peut-être obtenue par la minimisation du NLV/NLL, SCR/RSS, EQM/MSE ou REQM/RMSE.

EMV pour la régression linéaire homoscédastique

En ré-écrivant l'expression pour le SCR/RSS, nous obtenons:

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right)^2 = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 = (\mathbf{X}\boldsymbol{w} - \boldsymbol{y})^\top (\mathbf{X}\boldsymbol{w} - \boldsymbol{y})$$

Toujours sous l'optique du EMV (de la minimisation du risque empirique), nous voulons trouver une valeur pour $\boldsymbol{w}$ telle que $\nabla \mathrm{RSS}(\boldsymbol{w}) = \mathbf{0}$:

$$\hat{\boldsymbol{w}}_{\mathrm{mle}} \triangleq \operatorname*{argmin}_{\boldsymbol{w}} \mathrm{RSS}(\boldsymbol{w}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{y}$$

La solution $\hat{\boldsymbol{w}}_{\mathrm{mle}}$ porte aussi le nom de l'estimateur de la **méthode des moindres carrés ordinaire** (MCO/ordinary least squares (OLS)) en statistique.

Régularisation

L'utilisation de l'EMV (MLE) vs l'estimateur du maximum a posteriori (MAP) nous amène à développer le concept de **régularisation** dans le cadre de la minimisation du risque empirique. Pour mieux comprendre, prenons l'exemple de l'estimateur EMV pour la loi de Bernoulli pour un ensemble de données de 3 exemples, contenant seulement des observations de type "face". La valeur de l'EMV est alors:

$$\hat{\theta}_{\text{mle}} = N_1 / (N_0 + N_1) = 3/(3+0) = 1$$

Si nous simulons $\text{Ber}\left(y \mid \hat{\theta}_{\text{EMV}}\right)$, nous allons alors seulement observer des "faces", ce qui est peu probable intuitivement.

Régularisation

Dans l'exemple précédent, le modèle dispose de suffisamment de paramètres (capacité) pour parfaitement reproduire les données d'entraînement. C'est là un problème de généralisation. Une solution possible est de pénaliser la complexité du modèle pour mieux généraliser. L'expression pour le risque empirique devient alors:

$$\hat{R}(\boldsymbol{\theta}; \lambda) = \left[\frac{1}{N} \sum_{n=1}^{N} \ell\left(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n\right) \right] + \lambda C(\boldsymbol{\theta})$$

Régularisation et MAP

Si nous utilisons $C(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$, avec $p(\boldsymbol{\theta})$ étant un *a priori* pour $\boldsymbol{\theta}$, nous obtenons:

$$\hat{R}(\boldsymbol{\theta}; \lambda) = -\frac{1}{N} \sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}\right) - \lambda \log p(\boldsymbol{\theta})$$

Pour le choix de $\lambda = 1$, et en faisant abstraction du facteur $1/N$ nous avons:

$$\hat{R}(\boldsymbol{\theta}; \lambda) = -\left[\sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}\right) + \log p(\boldsymbol{\theta})\right] = -[\log p(\mathcal{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})] \ ,$$

puisque $\mathcal{D}$ est échantillonné i.i.d. La minimisation du risque empirique dans ce contexte coïncide alors avec l'estimateur du maximum a posteriori (MAP):

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{\theta} \mid \mathcal{D}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} [\log p(\mathcal{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \text{ constante }]$$

Statistique bayésienne

L'estimateur MAP provient de la statistique bayésienne qui s'intéresse à caractériser l'incertitude sur les paramètres. Cette connaissance de l'incertitude permet de réduire les chances de surapprentissage. À l'aide du théorème bayésien:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\mathcal{D} \mid \boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta})p(\mathcal{D} \mid \boldsymbol{\theta})}{\int p\left(\boldsymbol{\theta}'\right) p\left(\mathcal{D} \mid \boldsymbol{\theta}'\right) d\boldsymbol{\theta}'}$$

où $p(\boldsymbol{\theta} \mid \mathcal{D})$ est la **distribution a posteriori**, $p(\boldsymbol{\theta})$ la **distribution a priori** (prior distribution), $p(\mathcal{D} \mid \boldsymbol{\theta})$ la **fonction de vraisemblance** (likelihood function) et $p(\mathcal{D})$ la **vraisemblance marginale** (marginal likelihood). La **distribution prédictive a posteriori** (posterior predictive distribution) est ainsi $p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}$, qui peut-être vue comme une moyenne pondérée d'un nombre infini de modèles.

Exemple: MAP pour la loi de Bernoulli

Revenons à notre exemple précédent. Utilisons cette fois l'estimateur MAP et un a priori $p(\theta) = \mathrm{Beta}(\theta \mid a, b)$ pour lequel les valeurs $a, b > 1$ favorisent une solution pour θ s'approchant de $a/(a+b)$ L'a posteriori (qui est le log-vraisemblance plus le log a priori) est alors:

$$\begin{aligned}\ell(\theta) &= \log p(\mathcal{D} \mid \theta) + \log p(\theta) \\ &= [N_1 \log \theta + N_0 \log(1-\theta)] + [(a-1)\log(\theta) + (b-1)\log(1-\theta)]\end{aligned}$$

Le maximiseur pour ℓ (calculer la dérivée, résoudre pour zéro) est alors:

$$\theta_{\mathsf{map}} = \frac{N_1 + a - 1}{N_1 + N_0 + a + b - 2}$$

Exemple: MAP pour la loi de Bernoulli

Le choix de $a = b = 2$ a pour conséquence de favoriser les valeurs de θ proches de 0.5 et l'estimateur MAP est alors:

$$\theta_{\text{map}} = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

Ce choix particulier correspond à une technique bayésienne appelée "lissage de Laplace" (Laplace smoothing) ou "add-one smoothing" en anglais. Cette technique nous est particulièrement utile dans l'absence d'observation d'un type particulier (zero-count problem). Pour notre ensemble $\mathcal{D}$ contenant trois observations de "face" seulement, l'estimateur MAP prédit alors:

$$\theta_{\text{map}} = \frac{3 + 1}{3 + 0 + 2} = 0.8$$

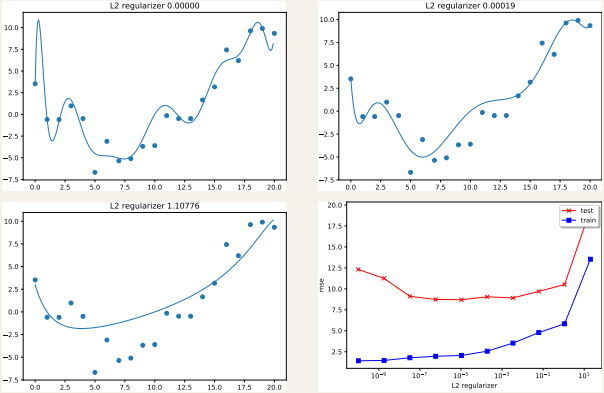
Régularisation et dégradation des pondérations

Prenons l'exemple de la régression polynomiale. Nous savons qu'il est facile de surapprendre dans ce contexte si nous choisissons une expansion polynomiale de degré trop élevée. Si nous choisissons alors un a priori gaussien $p(\boldsymbol{w})$ avec une moyenne de zéro, l'estimateur MAP devient alors:

$$\hat{\boldsymbol{w}}_{\text{map}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \operatorname{NLL}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2$$

avec $\|\boldsymbol{w}\|_2^2 = \sum_{d=1}^{D} w_d^2$. (Note: souvenez-vous que $\operatorname{NLL}(\boldsymbol{w})$, avec un modèle gaussien de vraisemblance, est équivalent à la minimisation de l'EQM). La pénalisation des paramètres avec la norme ℓ_2 est appelée la **dégradation des pondérations** (weight decay) ou simplement **régularisation** ℓ_2.

Régression ridge (par/de crêtes)



Dans un contexte de régression linéaire, l'utilisation de la dégradation des pondérations est appelée **régression ridge**. Cette figure montre le résultat de la régularisation ℓ_2 dans un cadre de régression polynomiale:

$$f(x; \boldsymbol{w}) = \sum^{D} w_d x^d = \boldsymbol{w}^\top \left[1, x, x^2, \ldots, x^D\right]$$

Choix de la valeur de régularisation

Une grande valeur du coefficient de régularisation λ met plus d'emphase sur l'importance de rester proche du a priori, alors qu'une petite valeur favorise les solutions qui minimise le risque empirique seulement.

De trop grande valeur de λ peuvent causer du **sous-apprentisage**, et du **surapprentissage** dans le cas contraire. Comment doit-on alors choisir λ? Par l'entremise d'un ensemble de validation $\mathcal{D}_{\text{valid}}$ (aussi appelé **ensemble de développement**).

Nous choisissons souvent de garder 80% des exemples pour l'entraînement, et 20% pour la validation.

Recherche par quadrillage

L'utilisation d'un ensemble d'entraînement nous permet d'implémenter un algorithme appelé **recherche par quadrillage** (grid search) pour **l'optimisation boîte noire pour la recherche d'hyperparamètres** (black box hyperparameter optimization). Pour y arriver, définissons d'abord le **risque empirique régularisé**:

$$\hat{R}_\lambda(\boldsymbol{\theta}, \mathcal{D}) \triangleq \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \ell(\boldsymbol{y}, f(\boldsymbol{x}; \boldsymbol{\theta})) + \lambda C(\boldsymbol{\theta})$$

Recherche par quadrillage

Pour un choix donné d'hyperparamètre λ, la solution optimale est dénotée:

$$\hat{\boldsymbol{\theta}}_\lambda\left(\mathcal{D}_{\text{train}}\right) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \hat{R}_\lambda\left(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}\right)$$

Nous définissons aussi le **risque de validation** comme étant:

$$\hat{R}_\lambda^{\text{val}} \triangleq \hat{R}_0\left(\hat{\boldsymbol{\theta}}_\lambda\left(\mathcal{D}_{\text{train}}\right), \mathcal{D}_{\text{valid}}\right)$$

Le risque de validation nous permet d'approximer correctement le risque (le "vrai" risque) aussi appelé **risque de la population** (population risk):

$$R(\boldsymbol{\theta}) \triangleq \mathbb{E}_{(x,y)\sim p(x)p(y|x)}\left[l(\boldsymbol{y}, f(\boldsymbol{x}))\right]$$

c'est-à-dire, l'espérance de la fonction de perte sous la distribution $(x, y) \sim p(x)p(y|x)$.

Recherche par quadrillage

Input: $\mathcal{S} \leftarrow \{\lambda_0, \ldots, \lambda_k\}$, un ensemble d'hyperparamètres

Output: λ^*, l'hyperparamètre optimal

$\mathcal{V} \leftarrow \emptyset$;

foreach $\lambda \in \mathcal{S}$ **do**

  $\hat{\boldsymbol{\theta}}_\lambda\left(\mathcal{D}_{\text{train}}\right) \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \hat{R}_\lambda\left(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}\right)$;

  $\mathcal{V} \leftarrow \mathcal{V} \cup \hat{R}_0\left(\hat{\boldsymbol{\theta}}_\lambda\left(\mathcal{D}_{\text{train}}\right), \mathcal{D}_{\text{valid}}\right)$;

end

$\lambda^* \leftarrow \underset{\lambda \in \mathcal{S}}{\operatorname{argmin}}\, \mathcal{V}$;

return λ^*;

Recherche par quadrillage

Avec la valeur λ^* obtenue grâce à l'algorithme prédécent, nous pouvons utiliser l'entièreté des données $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{valid}}$ pour obtenir finalement:

$$\hat{\boldsymbol{\theta}}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, R_{\lambda^*}(\boldsymbol{\theta}, \mathcal{D}).$$

Validation croisée

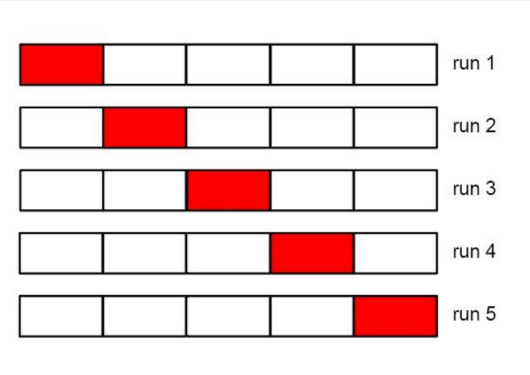
Lorsque nous avons trop peu de données d'entraînement pour définir un ensemble de validation, nous pouvons utiliser une technique alternative appelée **validation croisée** (cross-validation) pour choisir les hyperparamètres.

Pour y arriver, nous partitionnons les données en k **blocs** (folds), et de manière itérative, entraînons notre modèle sur tous les blocs sauf le k-ième, que nous réservons pour l'évaluation du modèle. Le **risque de validation croisée** est alors:

$$\hat{R}_\lambda^{\text{cv}} \triangleq \frac{1}{K} \sum_{k=1}^{K} \hat{R}_0 \left(\hat{\boldsymbol{\theta}}_\lambda \left(\mathcal{D}_{-k} \right), \mathcal{D}_k \right)$$

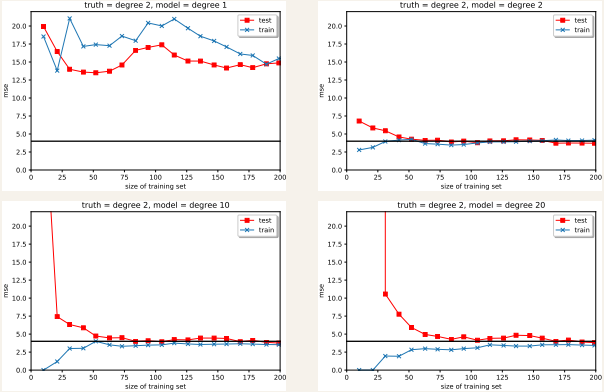
L'hyperparamètre retenu est alors $\lambda^* = \operatorname{argmin}_\lambda R_\lambda^{\text{cv}}$ et les paramètres du modèle sont estimés à nouveau sur l'entièreté des données $\hat{\boldsymbol{\theta}}^* = \operatorname{argmin}_{\boldsymbol{\theta}} R_{\lambda^*}(\boldsymbol{\theta}, \mathcal{D})$.

Validation croisée à $k = 5$ blocs



(Note: le cas particulier de $k = N$ est appelé **validation croisée tout sauf un** (leave-one-out cross validation).

Effet de la quantité de données



Pour un modèle de complexité fixe, l'augmentation de la taille de l'ensemble d'entraînement réduit les chances de surapprentissage.

Théorie de l'apprentissage statistique

Une autre approche pour quantifier les capacités de généralisation d'un modèle consiste à en dériver une borne supérieure grâce à la **théorie de l'apprentissage statistique** (statistical learning theory).

En classification binaire, si une telle borne est satisfaite, nous pouvons montrer que la minimisation du risque empirique résultera également en une faible valeur pour le risque de la population (probabilistiquement parlant). Nous disons alors que la **classe d'hypothèses** (hypothesis class) $\mathcal{H}$ est **PAC apprenable** (PAC learnable), un que le modèle $h \in \mathcal{H}$ trouvé satisfaisant cette borne est appelé **probably approximately correct** (PAC).

Borne de généralisation

Supposons que la classe d'hypothèse soit de dimension finie de taille $\dim(\mathcal{H}) = |\mathcal{H}|$.

Théorème 5.4.1 (PML): *Pour toute distribution sur les données p^* et tout ensemble $\mathcal{D}$ de taille $N_{\mathcal{D}}$ échantillonné à partir de p^*, la probabilité que l'erreur de généralisation d'un classifieur binaire dépasse ϵ est dans le pire cas est bornée par le haut par:*

$$P\left(\max_{h \in \mathcal{H}} |R(h) - \hat{R}(h, \mathcal{D})| > \epsilon\right) \leq 2 \dim(\mathcal{H}) e^{-2N_{\mathcal{D}}\epsilon^2}$$

où $\hat{R}(h, \mathcal{D}) = \frac{1}{N_{\mathcal{D}}} \sum_{i=1}^{N} \mathbb{I}\left(f(\boldsymbol{x}_i) \neq y_i^*\right)$ est le risque empirique et $R(h) = \mathbb{E}\left[\mathbb{I}\left(f(\boldsymbol{x}) \neq y^*\right)\right]$ est le risque de la population.

Preuve

Preuve. La preuve se base sur **l'inégalité d'Hoeffding** qui donne une borne supérieure sur la probabilité que la somme d'un nombre fini de variables aléatoires i.i.d. s'éloigne de leur moyenne dans une quantité donnée. Si $E_1, \ldots, E_{N_\mathcal{D}} \sim \mathrm{Ber}(\theta)$, alors pour tout $\epsilon > 0$,

$$P(|\bar{E} - \theta| > \epsilon) \leq 2e^{-2N_\mathcal{D}\epsilon^2}$$

où $\bar{E} = \frac{1}{N_\mathcal{D}} \sum_{i=1}^{N_\mathcal{D}} E_i$ est le taux d'erreur empirique et θ est le vrai taux d'erreur. Le second élément de la preuve se base sur **l'inégalité de Boole** (union bound) , selon lequel si $A_1, \ldots, A_d$ sont des ensembles d'événements, alors:

$$P\left(\cup_{i=1}^{d} A_i\right) \leq \sum_{i=1}^{d} P\left(A_i\right)$$

Preuve

Ces deux inégalités nous permettent de conclure que:

$$
\begin{aligned}
P\left(\max_{h\in\mathcal{H}} |R(h) - \hat{R}(h,\mathcal{D})| > \epsilon\right) &= P\left(\bigcup_{h\in\mathcal{H}} |R(h) - \hat{R}(h,\mathcal{D})| > \epsilon\right) \\
&\leq \sum_{h\in\mathcal{H}} P(|R(h) - \hat{R}(h,\mathcal{D})| > \epsilon) \\
&\leq \sum_{h\in\mathcal{H}} 2e^{-2N_{\mathcal{D}}\epsilon^2} = 2\dim(\mathcal{H})e^{-2N_{\mathcal{D}}\epsilon^2}
\end{aligned}
$$

Interprétation: la valeur de cette borne augmente pour une classe d'hypothèses plus grande (ie. $\dim(\mathcal{H})$) et diminue avec de plus grands ensembles d'entraînement (ie. $N_{\mathcal{D}}$).