COURS D'APPRENTISSAGE STATISTIQUE
(SYLVAIN ARLOT ET FRANCIS BACH)
MASTÈRE M2 PROBABILITÉS-STATISTIQUE 2011

COURS 3 - CONVEXIFICATION DU RISQUE

NOTES DE COURS PRISES PAR NELO MOLTER MAGALHÃES ET MAUD THOMAS

On cherche à optimiser le risque empirique. Cela n'est pas toujours possible, mais il existe des cas où on sait le faire, par exemple lorsque la perte est convexe.

1. Rappels d'optimisation

1.1. **Dualité Lagrangienne.** On cherche à minimiser sur $\mathcal{X} \subset \mathbb{R}^n, x \to f(x)$ telle que :

$$(*) \quad \begin{array}{ll} h_i(x) = 0, & i = 1, \dots, m \\ g_j(x) \leq 0, & j = 1, \dots, r \end{array}$$

avec $m + r = n$. On note $\mathcal{X}^*$ l'ensemble des $x \in \mathcal{X}$ vérifiant $(*)$.

Définition 1. On appelle Lagrangien du problème d'optimisation la fonction $\mathcal{L}$ définie par :

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x)$$

où $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+$
λ et μ sont appelés les multiplicateurs de Lagrange

Proposition 1.1.

$$\sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} \mathcal{L}(x, \lambda, \mu) = \begin{cases} f(x) & \textit{si } x \textit{ vérifie } (*) \\ \infty & \textit{sinon.} \end{cases}$$

Définition 2. On appelle fonction duale la fonction q définie par :

$$q(\lambda, \mu) = \inf_{x \in \mathcal{X}} \mathcal{L}(x, \lambda, \mu)$$

Proposition 1.2. *q est concave*

Remarque 1. La proposition ci-dessus est vraie sans aucune hypothèse.

Démonstration. $\mathcal{L}$ est affine en λ, μ, et l'infimum de fonctions linéaires est toujours concave. □

Proposition 1.3.

$$\forall \lambda, \mu, x \quad q(\lambda, \mu) \leq f(x)$$

Démonstration. Si x vérifie $(*)$, alors

$$q(\lambda, \mu) = \inf_{x \in \mathcal{X}} \mathcal{L}(x, \lambda, \mu) \leq \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^r_+} \mathcal{L}(x, \lambda, \mu) = f(x)$$

Sinon, l'inégalité est triviale. □

Date: Mercredi 2 Mars 2011.

1.2. **Problème dual : Maximiser $q(\lambda, \mu)$ en (λ, μ).**

Remarque 2. Par la proposition 1.3, on a :

$$\forall \lambda, \mu \quad q(\lambda, \mu) \leq \min_{x \in \mathcal{X}^*} f(x)$$

Donc :

$$\max_{(\lambda,\mu)} q(\lambda, \mu) \leq \min_{x \in \mathcal{X}^*} f(x)$$

Définition 3. On dit qu'il y a dualité faible si :

$$f^* := \min_{x \in \mathcal{X}^*} f(x) \geq q* := \max_{(\lambda,\mu)} g(\lambda, \mu)$$

Définition 4. On dit qu'il y a dualité forte si

$$f^* = q^*$$

Définition 5. On appelle conditions de Slater les hypothèses suivantes :
- h_i est une fonction affine pour tout i
- g_j est une fonction convexe pour tout j
- il existe $x \in \mathcal{X}$ tel que :

$$h_i(x) = 0, \quad \forall i = 1, \ldots, m$$
$$g_j(x) < 0, \quad \forall j = 1, \ldots, r$$

Proposition 1.4. *Si les hypothèses de Slater sont vérifiées alors il y a dualité forte.*

Remarque 3.
- Sous certaines hypothèses (le plus souvent satisfaites), le dual du dual est le primal
- Le dual n'est pas unique. En effet, par exemple, on a :

$$\min_{x:|x| \leq 1} f(x) = \min_{x:x^2 \leq 1} f(x)$$

On a deux fonctions g différentes donc deux problèmes duaux différents.

1.3. **Problèmes variationnels (ou minimax).** Il est rare que l'on ait pour une fonction K quelconque :

$$\min_x \max_\lambda K(x, \lambda) = \max_\lambda \min_x K(x, \lambda)$$

Mais cela est vrai si K est convexe par rapport à la variable x et concave par rapport à la variable λ.

1.4. **Condition de KKT (Karush - Kühn - Tucker).**

Rappel 1. Si f est de classe $\mathcal{C}^1$ et convexe alors une condition d'optimalité est : $\bigtriangledown f(x) = 0$

Supposons qu'il y ait dualité forte. On note :
- $x^* \in \text{argmin}_{x \in \mathcal{X}^*}\{f(x)\}$
- $(\lambda^*, \mu^*) \in \text{argmax}\{q(\lambda, \mu)\}$

On a alors :

$$\begin{aligned} f(x^*) =& q(\lambda^*, \mu^*) \\ =& \inf_x \{f(x) + \lambda^{*T} g(x) + \mu^{*T} h(x)\} \\ \leq& f(x^*) + \lambda^{*T} g(x^*) + \mu^{*T} h(x^*) \\ \leq& f(x^*) \end{aligned}$$

Donc toutes les inégalités sont en fait des égalités.
On obtient ainsi la première condition de KKT (complementary slackness)

$$\forall j, \lambda_j^* g_j(x^*) = 0$$

La deuxième condition de KKT est :

$$h(x^*) = 0 \text{ et } g(x*) \leq 0$$

Quand x vérifie la deuxième condition de KKT, on dit qu'il est admissible.

Et enfin la troisième condition de KKT :

$$x^* \in \operatorname{argmin} \mathcal{L}(x, \lambda, \mu)$$

Remarque 4. On peut montrer que ces conditions sont nécessaires et suffisantes pour avoir dualité forte.

Rappel 2. Soit $f : \mathbb{R}^d \to \mathbb{R}$ convexe.
Si on pose : $f^*(y) = \sup_{x \in \mathcal{X}} \{x^T y - f(x)\}$. Alors, $\forall x, f(x) = \sup_{y \in \mathbb{R}^d} \{x^T y - f^*(y)\}$

Remarque 5. f^* s'appelle la duale de Fenchel-Legendre de f.

2. Classification binaire (supervisée)

2.1. **Convexification du risque.** *Données* : $(X_i, Y_i)_{i=1,\dots,n} \in \mathcal{X} \times \{-1, 1\}$ indépendantes et identiquement distribuées.
But : trouver $f : \mathcal{X} \to \{-1, 1\}$ telle que :

$$\mathcal{R}(f) = E(\mathbf{1}_{f(X) \neq Y}) = P(f(X) \neq Y)$$

soit le plus petit possible.

Remarque 6. $\mathcal{R}(f)$ désigne le risque, ou encore la probabilité d'erreur de f.

Problème : $\{-1, 1\}$ n'est pas un espace vectoriel, donc $\{f : \mathcal{X} \to \{-1, 1\}\}$ non plus, ce qui peut poser une difficulté pour résoudre un problème de minimisation.

Nouveau but : trouver $f : \mathcal{X} \to \mathbb{R}$ et considérer alors la fonction : $x \to \operatorname{sign}(f(x))$ comme prédicteur, où :

$$\operatorname{sign}(a) = \begin{cases} 1 & \text{si } a > 0 \\ -1 & \text{si } a < 0 \\ 0 & \text{si } a = 0 \end{cases}$$

Risque : Le risque devient alors :

$$\mathcal{R}(f) = P(\operatorname{sign}(f(X)) \neq Y) = E(\mathbf{1}_{\operatorname{sign}(f(X)) \neq Y}) = E(\mathbf{1}_{Y f(X) \leq 0}) = E\Phi_{0-1}(Y f(X))$$

où : Φ_{0-1} est la perte $0-1$.
Risque empirique :

$$\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \Phi_{0-1}(Y_i f(X_i))$$

On cherche à minimiser le risque empirique, mais on ne peut pas le faire directement car Φ_{0-1} est ni continue, ni convexe.

Question : Quel lien y a-t-il entre la perte $0-1$ et les pertes convexes ?

2.2. **Exemples.**

Exemple 1. (1) Perte quadratique : ici $\Phi(u) = (u-1)^2$

$$\Phi(Y f(X)) = (Y - f(X))^2 = (f(X) - Y)^2$$

On retrouve les moindres carrés.

$$\mathcal{R}_\Phi(f) := E\Phi(Y f(X)) = E(f(X) - Y)^2$$

$\mathcal{R}_\Phi(f)$ est appelé le Φ-risque.

Rappel 3. On a vu précédemment : $\mathcal{R}(f) - \inf_f \{\mathcal{R}(f)\} \leq \sqrt{\mathcal{R}_\Phi(f) - \inf_f \{\mathcal{R}_\Phi(f)\}}$
Le membre de droite s'appelle l'excès de Φ-risque.

(2) Perte logistique : ici $\Phi(u) = \log(1 + e^{-u})$

$$\Phi(Yf(X)) = \log(1 + e^{-Yf(X)}) = -\log(\frac{1}{1 + e^{-Yf(X)}}) = -\log(\sigma(Yf(X))$$

où : $\sigma(v) = \frac{1}{1+e^{-v}}$ fonction sigmoïde.
Lien avec les probabilités : on considère le modèle défini par :

$$q(Y = 1|X = x) = \sigma(f(x)) \text{ et } q(Y = -1|X = x) = \sigma(-f(x))$$

Alors le risque est égal à $-$la log-vraisemblance conditionnelle : $E[-\log(q(Y|X))]$

(3) Perte hinge : ici $\Phi(u) = \max(1 - u, 0)$
Soit $f(x) = w^T x + b$, on appelle marge la quantité (ceci correspond à une interprétation géométrique) :

$$\frac{1}{||w||}$$

On cherchera donc à minimiser $||w||$. Un classifieur adapté à ce problème est appelé classifieur "maximum margin".

Remarque 7. Pour ce dernier exemple, il y a deux formulations possibles :
-La formulation SVM (=support vector machine) séparable :
$\min \frac{1}{2}||w||^2$ tel que $y_i(w^T x_i + b) \geq 1$
-La formulation SVM non séparable :
$\min \frac{1}{2}||w||^2 + c\sum_{i=1}^n \xi_i$ tel que

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

Cette dernière peut se reformuler comme suit : $\min \frac{1}{2}||w||^2 + c\sum_{i=1}^n \max(0, 1 - y_i(w^T x_i + b))$
Ou encore : $\min \frac{1}{2}||w||^2 + c\sum_{i=1}^n \Phi(y_i f(x_i)))$

Exemple 2. Si on note :

$$\mathcal{L}(w, \xi, \alpha, \beta) = \frac{1}{2}||w||^2 + c\sum_{i=1}^n \xi_i - \sum_{i=1}^n \beta_i \xi_i - \sum_{i=1}^n \alpha_i(y_i(w^T x_i + b) - 1 + \xi_i)$$

Les variables w, ξ sont liées au problème primal ; les variables α, β sont liées au problème dual.
Pour minimiser cette expression en w, en dérivant, on trouve : $0 = w - \sum_{i=1}^n \alpha_i y_i x_i$
En minimisant en ξ on trouve : $0 = c - \alpha_i - \beta_i$
En minimisant en b, on trouve : $0 = \sum_{i=1}^n \alpha_i y_i$

Et le problème dual s'écrit : $\max_\alpha \frac{-1}{2}||\sum_{i=1}^n \alpha_i y_i x_i||^2$ telle que :

$$0 = \sum_{i=1}^n \alpha_i y_i$$
$$0 \leq \alpha_i \leq c$$

Remarque 8. Les conditions de KKT s'écrivent :

$$(c - \alpha_i)\xi_i = 0, \qquad i = 1, \ldots, n$$
$$\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0, \quad i = 1, \ldots, n$$

On remarque alors que si $y_i(w^T x_i + b) - 1 > 0$, cela signifie qu'on est à droite du hinge et, comme $\xi_i = 0$, on a : $\alpha_i = 0$
De même, si $y_i(w^T x_i + b) - 1 = 0$, cela signifie qu'on est sur le hinge, et on a : $0 < \alpha_i < c$
Enfin, si $y_i(w^T x_i + b) - 1 < 0$, cela signifie qu'on est à gauche du hinge, et on a : $\alpha_i = c$
Notons que ceci est un cas particulier du théorème du représentant (cours suivant).

2.3. **Liens entre risque et Φ-risque.** Plaçons nous dans le cadre initial de ce paragraphe, avec une fonction $g : \mathcal{X} \to \{-1, 1\}$. Si on note $\eta(x) = P(Y = 1|X = x)$, alors le risque de g vérifie :

$$\mathcal{R}(g) = E\Phi_{0-1}(Yg(X)) = E[E(\mathbf{1}_{(g(X))\neq Y})|X = x)] \geq E\min(\eta(X), 1 - \eta(X))$$

Et le meilleur classifieur est $g^*(x) = \text{sign}(2\eta(x) - 1)$.

Lemma 2.1.

$$\mathcal{R}(g) - \mathcal{R}(g^*) = E[\mathbf{1}_{g(X)\neq g^*(X)}|2\eta(X) - 1|]$$

Démonstration.

$$\begin{aligned}
\mathcal{R}(g) - \mathcal{R}(g^*) &= E[E[\mathbf{1}_{g(x)\neq Y} - \mathbf{1}_{g^*(x)\neq Y}|X = x]] \\
&= E[\eta(X)(\mathbf{1}_{g(X)=-1} - \mathbf{1}_{g^*(X)=-1})] + E[(1 - \eta(X))(\mathbf{1}_{g(X)=1} - \mathbf{1}_{g^*(X)=1})] \\
&= E[(2\eta(X) - 1)(\mathbf{1}_{g(X)=-1} - \mathbf{1}_{g^*(X)=-1})] \\
&= E[(2\eta(X) - 1)(\mathbf{1}_{g(X)\neq g*(X)}\,\text{sign}(2\eta(X) - 1))]
\end{aligned}$$

D'où le résultat. □

Définition 6. Soit $f : \mathcal{X} \to \mathbb{R}$, on définit le Φ-risque conditionnel par

$$E[\Phi(Yf(X))|X = x]$$

Remarque 9. On a $E[\Phi(Yf(X))|X = x] = \eta(x)\Phi(f(x)) + (1 - \eta(x))\Phi(-f(x)) := C_{\eta(x)}(f(x))$
Avec $C_\eta(\alpha) = \eta\Phi(\alpha) + (1 - \eta)\Phi(-\alpha)$

Proposition 2.2. *Si on pose :* $H(\eta) = \inf_\alpha C_\eta(\alpha)$, *on a :*

$$\mathcal{R}^*_\Phi := \inf_g \mathcal{R}_\Phi(g) = E[H(\eta(X))]$$

Définition 7. Φ est calibrée si

$$\begin{array}{ll}
\textit{pour } \eta > \frac{1}{2}, & \inf_{\alpha<0} C_\eta(\alpha) > H(\eta) \\
\textit{pour } \eta < \frac{1}{2}, & \inf_{\alpha>0} C_\eta(\alpha) > H(\eta)
\end{array}$$

Proposition 2.3. *Soit* $\Phi : \mathcal{R} \to \mathbb{R}$ *convexe, on a :* Φ *calibrée* ⇔ Φ *dérivable en 0 et* $\Phi'(0) < 0$

Démonstration. Notons $h : \alpha \mapsto C_\eta(\alpha)$;
Comme Φ est convexe, h aussi :

$$\begin{array}{ll}
\text{h atteint son minimum sur} & \mathbb{R}_+ \Leftrightarrow h'_d(0) < 0 \\
\text{h atteint son minimum sur} & \mathbb{R}_- \Leftrightarrow h'_g(0) > 0
\end{array}$$

Par ailleurs, on a :

$$\begin{array}{l}
h'_d(0) = \eta\Phi'_d(0) - (1 - \eta)\Phi'_g(0) < 0 \text{ si } \eta > \frac{1}{2} \\
h'_g(0) = \eta\Phi'_g(0) - (1 - \eta)\Phi'_d(0) > 0 \text{ si } \eta < \frac{1}{2}
\end{array}$$

– ⇐ :
Comme Φ est dérivable en 0 : $\Phi'_d(0) = \Phi'_g(0)$,
donc, comme $\Phi'(0) < 0$:

$$\begin{array}{l}
h'(0) = \Phi'(0)(2\eta - 1) < 0 \text{ si } \eta > \frac{1}{2} \\
h'(0) = \Phi'(0)(2\eta - 1) > 0 \text{ si } \eta < \frac{1}{2}
\end{array}$$

Donc Φ est calibrée.

– ⇒ :
Si on a Φ calibrée, en faisant :

$$\begin{array}{l}
\eta \longrightarrow \frac{1}{2}^+, \text{ on a } : \frac{1}{2}\Phi'_d(0) - \frac{1}{2}\Phi'_g(0) \leq 0 \\
\eta \longrightarrow \frac{1}{2}^-, \text{ on a } : \frac{1}{2}\Phi'_g(0) - \frac{1}{2}\Phi'_d(0) \geq 0
\end{array}$$

Ainsi $\Phi'_d(0) \leq \Phi'_g(0)$; et comme Φ est convexe par hypothèse, on a $\Phi'_d \geq \Phi'_g$;

D'où le résultat. □

On suppose pour la suite Φ convexe, $\Phi'(0) < 0$ et Φ dérivable en 0.

Définition 8. On définit $\Psi(\theta) = \Phi(0) - H(\frac{1+\theta}{2})$

Ainsi $\Psi(\theta) = \Phi(0) - \inf_\alpha\{\frac{1+\theta}{2}\Phi(\alpha) + \frac{1-\theta}{2}\Phi(-\alpha)\}$

Proposition 2.4. *Ψ est paire, Ψ est convexe*
$\Psi \geq 0$
$\Psi(0) = 0 = \Phi(0) - \inf_\alpha \frac{\Phi(\alpha)+\Phi(-\alpha)}{2}$

Theorem 2.5.

$$\Psi(\mathcal{R}(f) - \mathcal{R}(f^*)) \leq \mathcal{R}_\Phi(f) - \mathcal{R}^*_\Phi$$

Démonstration. Par définition : $\Psi(\mathcal{R}(f) - \mathcal{R}(f^*)) = \Psi(E[\mathbf{1}_{\text{sign}(f(X))\neq g^*(X)}|2\eta(X) - 1|])$

$\leq E[\Psi(\mathbf{1}_{\text{sign}(f(X))\neq g^*(X)}|2\eta(X) - 1|)]$ (par Jensen)

$= E[\mathbf{1}_{\text{sign}(f(X))\neq g^*(X)}\Psi(2\eta(X) - 1))]$ (par parité de Ψ)

$= E[\mathbf{1}_{\text{sign}(f(X))\neq g^*(X)}(\Phi(0) - H(\eta(X)))]$ (par définition de Ψ)

On borne l'indicatrice par 1, $\Phi(0)$ par $\mathcal{R}_\Phi(f)$, et on conclut avec la Proposition 2.2.

□

Exemple 3.
(1) Perte quadratique
On trouve $H(\eta) = \inf_\alpha \eta\Phi(\alpha) + (1-\eta)\Phi(-\alpha) = \inf_\alpha \eta(\alpha-1)^2 + (1-\eta)(\alpha+1)^2$
Pour calculer l'expression de H, on optimise en α : en dérivant par rapport à α, on trouve,
$2\eta(\alpha-1) + 2(1-\eta)(\alpha+1) = 2\alpha - 2\eta + 1$
En annulant cette expression, on trouve $\alpha = \eta - \frac{1}{2}$
Ainsi, $H(\eta) = 1 - (2\eta-1)^2$; et $\Psi(\theta) = \theta^2$

(2) SVM
On trouve $H(\eta) = \inf_\alpha \eta\max(1-\alpha, 0) + (1-\eta)\max(1+\alpha, 0)$
En optimisant, on trouve : $H(\eta) = 2\min(\eta, 1-\eta)$, et $\Psi(\theta) = |\theta|$

(3) Perte logistique
On trouve par le même procédé : $\Phi(\theta) \geq \frac{\theta^2}{2}$

Références

[1] Convex Optimization, Boyd and Vandenberghe, Cambridge University Press, 2004
[2] Convex Analysis and Non Linear Optimization, Borwein and Lewis, Springer, 2006.