MAP569 Machine Learning II

PC2 : SVM, Enet

Reminder on KKT conditions

Let $f, -g_1, \dots, -g_n$ be $\mathcal{C}^1$ convex functions and define

$$\hat{x} = \underset{g_i(x) \geq 0}{\operatorname{argmin}} f(x)\ .$$

Karush-Kuhn-Tucker necessary conditions :
Define $L(x, \lambda) = f(x) - \sum_{i=1}^{n} \lambda_i g_i(x)$. Then, there exists $\hat{\lambda}$ such that

1. $\nabla_x L(\hat{x}, \hat{\lambda}) = 0\,;$
2. $\min(\hat{\lambda}_i, g_i(\hat{x})) = 0$ for $i = 1, \dots, n$.

Strong duality : in addition $\hat{\lambda} = \underset{\lambda \geq 0}{\operatorname{argsup}} \inf_x L(x, \lambda)$.

1 Support Vector Machine (SVM)

For any $w \in \mathbb{R}^p$, define the linear function $f_w(x) = \langle w, x \rangle$ from $\mathbb{R}^p$ to $\mathbb{R}$. For a given $R > 0$, we consider the set of linear functions $\mathcal{F} = \{f_w : \|w\| \leq R\}$. The aim of this exercise is to investigate the classifier $\widehat{h}_{\varphi,\mathcal{F}}(x) = \mathsf{sign}(\widehat{f}_{\varphi,\mathcal{F}}(x))$ where $\widehat{f}_{\varphi,\mathcal{F}}$ is solution to the convex optimisation problem

$$\widehat{f}_{\varphi,\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \varphi(-y_i f(x_i))\ ,$$

with $\varphi(x) = (1 + x)_+$ the *hinge* loss. The Lagrangian version of this minimization problem is

$$\widehat{f}_{\varphi,\mathcal{F}} = \underset{f_w \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f_w(x_i))_+ + \lambda \|w\|^2 \right\}\ ,$$

for some $\lambda > 0$.

1. Prove that $\widehat{f}_{\varphi,\mathcal{F}} = f_{\widehat{w}}$ where $\widehat{w}$ belongs to $V = \mathsf{Span}\{x_i : i = 1, \dots, n\}$.
2. Prove that $\widehat{w} = \sum_{j=1}^{n} \widehat{\beta}_j x_j$ where $\widehat{\beta} = [\widehat{\beta}_1, \dots, \widehat{\beta}_n]^T$ is solution to

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i (K\beta)_i)_+ + \lambda \beta^T K \beta \right\}\ ,$$

   with K the Gram matrix $K = [\langle x_i, x_j \rangle]_{1 \leq i,j \leq n}$.
3. Check that this minimization problem is equivalent to

$$\widehat{\beta} = \underset{\substack{\beta, \xi \in \mathbb{R}^n \text{ such that} \\ y_i (K\beta)_i \geq 1 - \xi_i \\ \xi_i \geq 0}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \beta^T K \beta \right\}.$$

4. From the KKT conditions, check that $\widehat{\beta}_i = y_i\widehat{\alpha}_i/(2\lambda)$, for $i = 1,\dots,n$ with $\widehat{\alpha}_i$ fulfilling $\min(\widehat{\alpha}_i, y_i(K\widehat{\beta})_i - (1-\widehat{\xi}_i)) = 0$ et $\min(1/n - \widehat{\alpha}_i, \widehat{\xi}_i) = 0$.
5. Prove the following properties
   — if $y_i\widehat{f}_{\varphi,\mathcal{F}}(x_i) > 1$ then $\widehat{\beta}_i = 0$;
   — if $y_i\widehat{f}_{\varphi,\mathcal{F}}(x_i) < 1$ then $\widehat{\beta}_i = y_i/(2\lambda n)$;
   — if $y_i\widehat{f}_{\varphi,\mathcal{F}}(x_i) = 1$ then $0 \le \widehat{\beta}_i y_i \le 1/(2\lambda n)$.
6. Give a geometric interpretation of this result.
7. From the strong duality, prove that $\widehat{\alpha}_i$ is solution to the dual problem

$$\widehat{\alpha} = \underset{0\le\alpha_i\le 1/n}{\operatorname{argmax}} \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{4\lambda}\sum_{i,j=1}^n K_{i,j} y_i y_j \alpha_i \alpha_j \right\}.$$

2 Elastic-Net

The Elastic-Net estimator involves both a ℓ^2 and a ℓ^1 penalty. It is meant to improve the Lasso estimator when the columns of $\mathbf{X}$ are strongly correlated. It is defined for $\lambda, \mu \ge 0$ by

$$\widehat{\beta}_{\lambda,\mu} \in \underset{\beta\in\mathbb{R}^p}{\mathsf{argmin}}\, \mathcal{L}(\beta) \quad \text{with} \quad \mathcal{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2 + \mu|\beta|_{\ell^1}.$$

In the following, we assume that the columns of $\mathbf{X}$ have norm 1.

1. Check that the partial derivative of $\mathcal{L}$ with respect to $\beta_j \neq 0$ is given by

$$\partial_j \mathcal{L}(\beta) = 2\left((1+\lambda)\beta_j - R_j + \frac{\mu}{2}\operatorname{sign}(\beta_j)\right) \quad \text{with} \quad R_j = \mathbf{X}_j^T\Big(Y - \sum_{k:\, k\neq j} \beta_k \mathbf{X}_k\Big).$$

2. Prove that the minimum of $\beta_j \to \mathcal{L}(\beta_1,\dots,\beta_j,\dots,\beta_p)$ is reached at

$$\beta_j = \frac{R_j}{1+\lambda}\left(1 - \frac{\mu}{2|R_j|}\right)_+.$$

3. Propose an algorithm to compute the Elastic-Net estimator.

The Elastic-Net procedure is implemented in the `R` package `glmnet` available at
`http://cran.r-project.org/web/packages/glmnet/`.