Introduction to Machine Learning

Support Vector Machines, Elastic-net

Reminder on KKT conditions

Let \( f, -g_1, \ldots, -g_n \) be \( C^1 \) convex functions and define

\[
\hat{x} = \arg\min_{g_i(x) \geq 0} f(x).
\]

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 \( i = 1, \ldots, n \).

Strong duality: in addition \( \hat{\lambda} = \arg\sup_{\lambda \geq 0} \inf_x L(x, \lambda) \).

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 \( \hat{h}_\varphi, F(x) = \text{sign}(\hat{f}_\varphi, F(x)) \) where \( \hat{f}_\varphi, F \) is solution to the convex optimisation problem

\[
\hat{f}_\varphi, F = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varphi(-y_i f(x_i)) + \lambda \|w\|^2,
\]

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

\[
\tilde{f}_\varphi, F = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f_w(x_i))_+ + \lambda \|w\|^2 \right\},
\]

for some \( \lambda > 0 \).

1. Prove that \( \tilde{f}_\varphi, F = f_{\tilde{w}} \) where \( \tilde{w} \) belongs to \( V = \text{Span}\{x_i : i = 1, \ldots, n\} \).
2. Prove that \( \tilde{w} = \sum_{j=1}^{n} \tilde{\beta}_j x_j \) where \( \tilde{\beta} = [\tilde{\beta}_1, \ldots, \tilde{\beta}_n]^T \) is solution to

\[
\tilde{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i (K\beta)_i)_+ + \lambda \beta^T K \beta \right\},
\]

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

\[
\tilde{\beta} = \arg\min_{\beta, \xi \in \mathbb{R}^n \text{ such that } y_i (K\beta)_i \geq 1 - \xi_i, \xi_i \geq 0} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \tilde{\beta}^T K \tilde{\beta} \right\}.
\]
4. From the KKT conditions, check that $\hat{\beta}_i = y_i/\hat{\alpha}_i/(2\lambda)$, for $i = 1, \ldots, n$ with $\hat{\alpha}_i$ fulfilling $\min(\hat{\alpha}_i, y_i(K\hat{\beta})_i - (1 - \hat{\xi}_i)) = 0$ et $\min(1/n - \hat{\alpha}_i, \hat{\xi}_i) = 0$.

5. Prove the following properties
   — if $y_i f_{\hat{\phi}, F}(x_i) > 1$ then $\hat{\beta}_i = 0$;
   — if $y_i f_{\hat{\phi}, F}(x_i) < 1$ then $\hat{\beta}_i = y_i/(2\lambda n)$;
   — if $y_i f_{\hat{\phi}, F}(x_i) = 1$ then $0 \leq \hat{\beta}_i y_i \leq 1/(2\lambda n)$.

6. Give a geometric interpretation of this result.

7. From the strong duality, prove that $\hat{\alpha}_i$ is solution to the dual problem

$$\hat{\alpha} = \arg\max_{0 \leq \alpha_i \leq 1/n} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{4\lambda} \sum_{i,j=1}^{n} K_{i,j} y_i y_j \alpha_i \alpha_j \right\}.$$ 

2 Elastic-Net

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

$$\hat{\beta}_{\lambda,\mu} \in \arg\min_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta) \quad \text{with} \quad \mathcal{L}(\beta) = \|Y - X\beta\|^2 + \lambda \|\beta\|^2 + \mu |\beta|_{\ell^1}.$$ 

In the following, we assume that the columns of $X$ have norm 1.

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

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

2. Prove that the minimum of $\beta_j \to \mathcal{L}(\beta_1, \ldots, \beta_j, \ldots, \beta_p)$ 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/.