Exercise no 4

Corentin Lacroix 1812554

MTH 6312: Méthodes Statistiques D'Apprentissage

October 8, 2015

Exercise 1 Show the kernel density estimator $\hat{f}(y) = \frac{1}{n\lambda}\sum_{i=1}^{n} K(\frac{y_i - y}{\lambda})$ for any non-negative kernel that $\int_{-\infty}^{\infty} K(y)dy = 1$ is a probability density.

Hint : you must show $\hat{f}(y) \geq 0$ and $\int_{-\infty}^{\infty} \hat{f}(y)dy = 1$.

Solution 1

(i) As $\forall x \in \mathcal{R}, K(x) \geq 0$, we have trivially $\forall y \in \mathcal{R}, \hat{f}(y) = \frac{1}{n\lambda}\sum_{i=1}^{n} K(\frac{y_i - y}{\lambda}) \geq 0$

(ii) Let's try to compute $\int_{-\infty}^{+\infty} f(y)dy$ (we should first try to simplify the integral on a non infinite interval $[-A, A], A \geq 0$, -notably for inversing sum and integral or applying a variable change- and then find the limit for $A \to \infty$ of the simplified quantity but here calculation is simple so we can pass this step)

$\int_{-\infty}^{+\infty} f(y)dy = \frac{1}{n\lambda}\int_{-\infty}^{+\infty}\sum_{i=1}^{n} K(\frac{y_i - y}{\lambda})dy = \frac{1}{n\lambda}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K(\frac{y_i - y}{\lambda})dy$

Then with the change $u_i = \frac{y_i - y}{\lambda}, du_i = -\frac{dy}{\lambda}\lambda > 0$ applied to our n integrals, we get :

$$\int_{-\infty}^{+\infty} f(y)dy = -\frac{1}{n\lambda}\sum_{i=1}^{n}\lambda\int_{+\infty}^{-\infty} K(u)du = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K(u)du = 1 because \int_{+\infty}^{-\infty} K(u)du = 1$$

We can conclude that

$$\boxed{f \text{ is a probability density}}$$

Exercise 2 Show that the eigen values of $\mathbf{A}+\lambda\mathbf{I}$ equals $\lambda_i+\lambda$ where λ_i's are the eigenvalues of $\mathbf{A}$. Use this result to argue where the ridge regression is useful.

Solution 2 Let A be a matrix of $\mathcal{R}^n$ and $\lambda \in \mathcal{R}$

$$\begin{aligned} \lambda_i \; eigen \; value \; of \; A &\iff \exists x_i \in \mathcal{R}^n, Ax_i = \lambda_i x_i \\ &\iff \exists x_i \in \mathcal{R}^n, (A + \lambda_i I)x_i = (\lambda_i + \lambda)x_i \\ &\iff \lambda_i + \lambda \; eigen \; value \; of \; A + \lambda I \end{aligned} \tag{1}$$

Moreover, we know that a $n \times n$ symetric matrix ($X^T X$ or $X^T X + \lambda I$ for example) always admits n different eigen vectors μ_i associated to real eigen values λ_i. So thanks to this and the previous results, we can deduce that if $\lambda > 0$ all eigen values of $X^T X + \lambda I, (\lambda_i + \lambda)$ are non zero eigen values (because are equals to at least λ. And thus, $X^T X + \lambda I$ is inversible and there is one unique solution to the Ridge linear regression Least Square problem $\hat{\beta}_{ridge}$.

Exercise 3 Show the degrees of freedom of the ridge regression $\mathrm{df}_\lambda = \mathrm{tr}\{X(X^T X + \lambda I)^{-1} X^T\} = \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + \lambda}$ where λ_j is the eigenvalues of $X^T\ X$.

Hint: use the singular value decomposition of X.

Solution 3 The singular value decomposition of a $n \times p$ matrix says :

$\forall M \;\; n \times p$ with real coefficients matrix, $\exists \; Q \; n \times n$ orthogonal matrix, $P \; p \times p$ orthogonal matrix and $D \; p \; \times n$ diagonal per block matrix s.t $M = QDP^T$

And then :

$$
\begin{aligned}
X^T(X^TX+\lambda I)^{-1})X^T &= X((QDP^T)^TQDP^T+\lambda I)^{-1}X^T \\
&= X(PD^TQ^TQDP^T+\lambda I)^{-1}X^T \\
&= X(PD^TDP^T+\lambda I)X^T \\
&= X(P(D^TD+\lambda I)P^T)^{-1}X^T \\
&= QDP^TP(D^TD+\lambda I)^{-1}P^TX^T \\
&= QD(D^TD+\lambda I)^-1D^TQ^T
\end{aligned}
\tag{2}
$$

But D^TD is diagonal and has the same eigen values than X^TX (we can prove it by writing X^TX). So with $D^TD = (\lambda_i)_{i\in[1,n]}$ (this also means non zero pseudo diagonal values in D will be $\sqrt{\lambda_i)}$, we have : $X^T(X^TX+\lambda I)^{-1})X^T = QDdiag(\frac{1}{\lambda_i+\lambda})D^TQ^T$

Then, $Ddiag(\frac{1}{\lambda_i+\lambda})D$ is also diagonal with an ensemble of diagonal values being smaller or larger than the diagonal values ensemble of D^TD (it depends in fact on wheter n or p is smaller). Diagonal values of $Ddiag(\frac{1}{\lambda_i+\lambda})D$ are $\{\frac{\lambda_i}{\lambda_i+\lambda}\}$ with more or less elements : if $p < n$ all non zero diagonal values are removed from this ensemble ; if $p > n$, zero values are added to this ensemble containing only non zero values. As only zero values are added/removed to the diagonal values of $Ddiag(\frac{\lambda_i}{\lambda_i})D$, this doesn't change its trace.

Moreover, $Tr(QD(D^TD+\lambda I)D^TQ^T) = Tr(D(D^TD+\lambda I)D^T)$ (left and right multiplication of a matrix by a matrix and its inverse.

Consequently we have

$$\boxed{Tr(X^T(X^TX+\lambda I)^{-1})X^T) = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i+\lambda}}$$

So

Exercise 4 Find the maximum likelihood estimator of $\boldsymbol{\beta}$ for the weighted linear regression. Weighted linear regression or *General Linear Model* is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ while $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W})$ and $\mathbf{W}$ is the known variance covariance matrix of $\boldsymbol{\varepsilon}$. A general linear model is called ordinary linear regression if $\mathbf{W} = \sigma^2\mathbf{I}$ for a known σ^2.

Solution 4 The random variable $Y|X \sim N(X\beta, W)$ with X fixed. With this density function (and the realization Y of this random variable) we can express the likelihood of β given a realization Y of $\boldsymbol{Y}|X$:

And $L(\beta) = f(\beta) = \frac{1}{|W|^{\frac{1}{2}}\sqrt{2\pi}} \exp^{-\frac{1}{2}(Y-X\beta)^T W^{-1}(Y-X\beta)}$
The log likelihood $l(\beta)$ is $l(\beta) = -\frac{1}{2}ln(|W|) - \frac{1}{2}ln(2\pi) - \frac{1}{2}((Y-X\beta)^T W^{-1}(Y-X\beta)$

And the log likelihood estimator given by $\hat{\beta} = argmax_\beta(l(\beta))$ is given by the equation $\frac{\partial l}{\partial \beta}(\hat{\beta}) = 0$ (convex optimization problem in coefficients of β).

With :

$$\begin{aligned}
\frac{\partial l}{\partial \beta}(\beta) &= -\frac{1}{2}(\frac{\partial (Y-X\beta)^T}{\partial \beta}W^{-1}(Y-X\beta) + (Y-X\beta)^T W^{-1}(\frac{\partial (Y-X\beta)}{\partial \beta} \\
&= -\frac{1}{2}(-X^T W^{-1}(Y-X\beta) - (Y-X\beta)^T W^{-1} X) \\
&= X^T W^{-1}(Y-X\beta) \textit{ by taking the transpose of the second member } (1\times 1 \textit{ matrix and W dia}
\end{aligned} \tag{3}$$

Finally, $\hat{\beta}$ is given by :

$$\boxed{X^T W^{-1}(Y-X\hat{\beta}) = 0 \iff \hat{\beta} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y \textit{ under assumption that } X^T W^{-1} X \textit{ is invers}}$$

Exercise 5 How do you compute the coefficients of the weighted linear regression? Write the steps of the computations.

Solution 5 The problem comes from inversing $X^T W^{-1} X$, a numerical computational process that can give very instable results.

Given the equation $X^T W^{-1} X\hat{\beta} = X^T W^{-1} Y$ that is a linear system, we can compute $\hat{\beta}$ by :

(i) Computing W^{-1} which is easy because W is diagonal

(ii) Computing $X^T W^{-1} X$ and $X^T W{-}1Y$ that is succession of + and x

(iii) solving the linear p unknown variables system $X^T W^{-1} X\hat{\beta} = X^T W^{-1} Y$

In R, the code would be :

```
> Winv <- solve(W)

> beta_opt <- solve(t(X)%*%Winv%*%X, t(X)%*%Winv%*%Y)
```

Exercise 6 How do you fit a weighted linear regression using a code that only fits the ordinary linear regression?

Solution 6 By $Y = X\beta + \varepsilon$ with $\varepsilon \sim N(0, W)$ and multiplying by the left the 2 sides of the equation by $W^{-\frac{1}{2}}$, we get :

$W^{-\frac{1}{2}}Y = W^{-\frac{1}{2}}X\beta + W^{-\frac{1}{2}}\varepsilon \iff Y' = X'\beta + \varepsilon'$ with $\varepsilon' \sim N(W^{-\frac{1}{2}}0, W^{-\frac{1}{2}}WW^{-\frac{1}{2}}) \sim N(0, I)$

And the soluion of ordinary linear regression problem is for Y', X' is :

$$\begin{aligned}
\hat{\beta} &= (X'^{T}X)X^{T}Y \\
&= ((W^{-\frac{1}{2}}X)^{T}(W^{-\frac{1}{2}}X))^{-1}(W^{-\frac{1}{2}}X)^{T}(W^{-\frac{1}{2}}X)) \\
&= (X^{T}W^{-1}X)^{-1}X^{T}W^{-1}Y
\end{aligned} \tag{4}$$

This is also the solution os the weighted linear regression for Y and X. To solve the weighted linear regression problem, we just have to get the solution os the ordinary linear regression for Y' and X'.

Exercise 7 Show that the kernel smoothing (weighted average) is the solution of the following optimization if $f_\theta(x) = \theta_0$

$$\hat{\theta}(x_0) = \text{argmin}_\theta \sum_{i=1}^{N} K(x_0, x_i)\{y_i - f_\theta(x_i)\}^2,$$

$$\hat{f}(x_0) = f_{\hat{\theta}}(x_0)$$

Solution 7 If $f_\theta(x) = \theta$ the optimization problem becomes

$$\hat{\theta}(x_0) = \text{argmin}_\theta \sum_{i=1}^{N} K(x_0, x_i)\{y_i - \theta\}^2 \ (1),$$

$$\hat{f}(x_0) = f_{\hat{\theta}(x_0)}(x_0) = \hat{\theta}(x_0) \ (2)$$

With $\varphi(\theta) = \sum_{i=1}^{n} K(x_0, x_i)(y_i - \theta)^2$, minimizing φ is equivalent to solving :

$$\begin{aligned} \frac{\partial \varphi}{\partial \beta}(\hat{\beta}) = 0 &\iff -2\sum_{i=1}^{n} K(x_0, x_i)(y_i - \hat{\theta}) = 0 \\ &\iff \hat{\theta} = \frac{1}{\sum_{i=1}^{n} K(x_0, x_i)} \sum_{i=1}^{n} K(x_0, x_i) y_i \end{aligned} \tag{5}$$

So $\hat{f}$ is also the solution of the weigthed average problem.

Exercise 8 Find the link between this optimization problem and the weighted linear regression.

Solution 8

$$\begin{aligned} \hat{\beta} \ solution \ of \ weighted \ linear \ regression &\iff \hat{\beta} = argmax_{\beta}(-\frac{1}{2}((Y - X\beta)^T W^{-1}(Y - X\beta)) \\ &\iff \hat{\beta} = argmin_{\beta}(\sum_{j=1}^{n} \frac{1}{\sigma_j^2}(y_j - \sum_{i=1}^{n} x_{ji}\beta_i)^2) \end{aligned} \tag{6}$$

Consequently here, with weights $W(x_0, x_i) = W(x_i) = W(i) = \frac{1}{\sigma_i^2}$ and the function $f_\beta(x_0) = \sum_{j=1}^{n} \beta_i x_{0i}$, we have an optimization problem of the same class than previously but with global weights depending only on the ith inputs x_i. Weighted linear regression is a kernel smoothing method (but with neighborhoods being global and weights not depending on the new input x_0).

Exercise 9 Find the solution of $\hat{f}$ for $f_\theta(x) = \theta_0 + \theta_1 x$?

Solution 9 The optimisation problem is :

$$\hat{\theta}(x_0) = \text{argmin}_\theta \sum_{i=1}^{N} K(x_0, x_i)\{y_i - \theta_0 - \theta_1 x_i\}^2 \ (1),$$

$$\hat{f}(x_0) = f_{\hat{\theta}(x_0)}(x_0) = \hat{\theta}_0(x_0) + \hat{\theta}_1(x_0)x_1 \ (2)$$

Let φ be : $\forall \theta \in \mathcal{R}, \varphi(\theta) = \sum_{i=1}^{N} K(x_0, x_i)(y_i - \theta_0 - \theta_1 x_i)^2$

$$\hat{\theta}(x_0) = \text{argmin}_\theta \sum_{i=1}^{N} K(x_0, x_i)\{y_i - \theta_0 - \theta_1 x_i\}^2 \Longleftrightarrow \frac{\partial \varphi}{\partial \theta}(\hat{\theta}) = 0 \tag{7}$$

Then $\frac{\partial \varphi}{\partial \theta_0}(\hat{\theta}) \Longleftrightarrow \hat{\theta}_0 = \frac{1}{\sum_{i=1}^{N} K(x_0,x_i)} \sum_{i=1}^{N} K(x_0, x_i)(y_i - \hat{\theta}1 x_i) = \overset{+}{y} - \theta_1 \overset{+}{x}$ with $\overset{+}{y}$, $\overset{+}{x}$ being weighted means of $y = (y_1, \ldots, y_N)$ and $x = (x_1, \ldots, x_N)$

And

$$\begin{aligned} \frac{\partial \varphi}{\partial \theta_1}(\hat{\theta}) = 0 &\Longleftrightarrow \sum_{i=1}^{N} K(x_0, x_i)x_i(y_i - \hat{\theta}1 x_i) = 0 \\ &\Longleftrightarrow \sum_{i=1}^{N} K(x_0, x_i)x_i \hat{\theta}_1(\overset{+}{x} - x_i) = \sum_{i=1}^{N} K(x_0, x_i)x_i(y_i - \overset{+}{y}) \qquad (8) \\ &\Longleftrightarrow \hat{\theta}_1 = \frac{\sum_{i=1}^{N} K(x_0, x_i)x_i(y_i - \overset{+}{y})}{\sum_{i=1}^{N} K(x_0, x_i)x_i(\overset{+}{x} - x_i)} \end{aligned}$$

With $\sum_{i=1}^{N} K(x_0, x_i)x_i \overset{+}{y} = \frac{1}{\sum_{i=1}^{N} K(x_0,x_i)} \sum_{j=1}^{N} K(x_0, x_j)y_j \times \sum_{i=1}^{N} K(x_0, x_i)x_i = \overset{+}{x} \sum_{j=1}^{N} K(x_0, x_j)y_j = \sum_{i=1}^{N} K(x_0, x_i) \times \overset{+}{y}\overset{+}{x}$,

we have $\sum_{i=1}^{N} K(x_0, x_i)x_i(y_i - \overset{+}{y}) = \sum_{i=1}^{N} K(x_0, x_i)(y_i x_i - \overset{+}{y} x_i - \overset{+}{x} y_i + \overset{+}{y}\overset{+}{x}) = \sum_{i=1}^{N} K(x_0, x_i)(x_i - \overset{+}{x})(y_i - \overset{+}{y})$

And with a similar trick for the denominator, we easily show that :

$$\boxed{\hat{\theta}_1 = \frac{\sum_{i=1}^{N} K(x_0, x_i)(x_i - \overset{+}{x})(y_i - \overset{+}{y})}{\sum_{i=1}^{N} K(x_0, x_i)(x_i - \overset{+}{x})^2}}$$

Exercise 10 Find the solution of $\hat{f}$ for $f_\theta(x) = \theta_0 + \sum_{j=1}^{M} \theta_j x^j$?

Solution 10 Solution not found...

Exercise 11 Find the linearly constrained least squares estimator $(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$ subject to $\beta = \mathbf{b}$ in which $\mathbf{T}$ and $\mathbf{b}$ both are known. How do you compute this estimator efficiently?

Hint: use the Lagrangian dual.

Solution 11 We want to minimize $(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$ subject to $\beta = \mathbf{b}$

By using Lagrangian dual optimization method, this is equivalent to minimize $\varphi(\beta) = (Y - X\beta)^T(Y - X\beta) + \mu T\beta$ with $\mu > 0$. This is a convex optimization problem (toward β), thus minimum of the function is given by :

$$
\begin{aligned}
\frac{\partial\varphi}{\partial\beta}(\hat{\beta}) = 0 &\iff -2X^T(Y - X\beta) + \mu T = 0 \\
&\iff 2X^T X\beta = 2X^T Y + \mu T \\
&\iff \hat{\beta} = (X^T X)^{-1}(X^T Y + \frac{1}{2}\mu T)
\end{aligned}
\tag{9}
$$

Thus, the linearly constrained least squares estimator is

$$\boxed{\hat{\beta} = (X^T X)^{-1}(X^T Y + \frac{1}{2}\mu T)}$$