Optimal models

Fabrice Rossi

In the following exercises, if l is loss function and $\mathcal{D}$ a data set on $\mathcal{X} \times \mathcal{Y}$, $\widehat{y}_l^*(\mathcal{D})$ is the best value according to the empirical risk based on l for a constant model, i.e.

$$\widehat{y}_l^*(\mathcal{D}) = \arg\min_{p \in \mathcal{Y}} \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},y) \in \mathcal{D}} l(p,y)$$

Exercise 1

Prove that if $\mathcal{Y} = \mathbb{R}$ and $l_2(p,t) = (p-t)^2$, then $\widehat{y}_{l_2}^*(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},y) \in \mathcal{D}} y$. Hint: study the derivative of $R(p) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},y) \in \mathcal{D}} l_2(p,y)$.

Exercise 2

In this exercise, we assume that $\mathcal{Y} = \{-1,1\}$ and that $l_0(p,t) = \mathbf{1}_{p \neq t}$.

Question 1 Prove that $\widehat{y}_{l_0}^*(\mathcal{D})$ is the most frequent value of y in $\mathcal{D}$.

Question 2 Can the previous result be extended to arbitrary finite $\mathcal{Y}$? What may happen more frequently when $|\mathcal{Y}|$ is large than when $|\mathcal{Y}|$ is small (for a given data set size)?

Let l_λ be the loss function defined by the following table

$l_\lambda(p,t)$		t	
		-1	1
p	-1	0	λ
	1	1	0

where $\lambda > 0$.

Question 3 Provide a simple characterisation of $\widehat{y}_{l_\lambda}^*(\mathcal{D})$.

Exercise 3

In this exercise, we assume that $\mathcal{Y} = \mathbb{R}$, $|\mathcal{D}| = 2m+1$ and all y from the data set are distinct (with $m \geq 1$). According, we sort them into $y_1 < y_2 < \ldots < y_{2m+1}$. We use the absolute loss function, i.e. $l_1(p,t) = |p-t|$. The objective of the exercise is to prove that $\widehat{y}_{l_1}^*(\mathcal{D}) = y_{m+1}$, that is the median of the $(y_k)_{1 \leq k \leq 2m+1}$.

We define $R(p) = \sum_{k=1}^{2m+1} |p - y_k|$. It is obvious that any minimizer of R is equal to $\widehat{y}_{l_1}^*(\mathcal{D})$.

Question 1 Give formulas with no absolute value for $R(p)$ when $p \leq y_1$ and when $p \geq y_{2m+1}$.

Question 2 Prove that $R(p)$ is decreasing on the interval $[y_1; y_2]$.

Question 3 Give a general formula without absolute value for $R(p)$ when $p \in [y_l,y_{l+1}]$, as well as on $]-\infty,y_1]$ and $[y_{2m+1},\infty[$.

Question 4 Show that R is monotonous on intervals of the form $[y_l,y_{l+1}]$, and on $]-\infty,y_1]$ and $[y_{2m+1},\infty[$.

Question 5 Use the previous results to conclude.

Question 6 Study the case when $|\mathcal{D}| = 2m$, for $m \geq 1$ with distinct values for all the y).