Universidad Técnica Federico Santa María
Departamento de Informática

Pauta Control 1 - Máquinas de Aprendizaje Computacional (INF-578)

Semestre II 2017 - Viernes 13.10.17

Respuestas:

1. Falso, ya que al aumentar la cardinalidad del espacio de hipótesis, hay menor garantía de haber encontrado la hipótesis óptima que modele la distribución conjunta subyacente usando el conjunto de entrenamiento. Esto ocurre, porque en un espacio más grande se puede encontrar una mayor cantidad de hipótesis con error 0 en el conjunto de entrenamiento.

2. $$\frac{1}{2}(\boldsymbol{Y}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})^T(\boldsymbol{Y}-\tilde{\boldsymbol{X}}\boldsymbol{\beta}) = \frac{1}{2}\left(\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta}\right)$$
$$\begin{aligned}\frac{1}{2}(\boldsymbol{Y}-\tilde{\boldsymbol{X}}\boldsymbol{\beta})^T(\boldsymbol{Y}-\tilde{\boldsymbol{X}}\boldsymbol{\beta}) &= \frac{1}{2}\left(\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T(\boldsymbol{X}+\boldsymbol{E})\boldsymbol{\beta} + \boldsymbol{\beta}^T(\boldsymbol{X}+\boldsymbol{E})^T(\boldsymbol{X}+\boldsymbol{E})\boldsymbol{\beta}\right)\\ &= \frac{1}{2}\left(\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{Y}^T\boldsymbol{E}\boldsymbol{\beta} + \boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{E}^T\boldsymbol{X} + \boldsymbol{X}^T\boldsymbol{E} + \boldsymbol{E}^T\boldsymbol{E})\boldsymbol{\beta}\right)\\ &= \frac{1}{2}\left(\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{Y}^T\boldsymbol{E}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^T\boldsymbol{E}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{\beta}\right)\end{aligned}$$

   Entonces, por la linealidad del valor esperado,

$$\begin{aligned}\mathbb{E}_\epsilon\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\beta}\tilde{\boldsymbol{X}})^T(\boldsymbol{Y}-\boldsymbol{\beta}\tilde{\boldsymbol{X}}) &= \frac{1}{2}(\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{\beta}\boldsymbol{X} - 2\boldsymbol{Y}^T\boldsymbol{\beta}\mathbb{E}_\epsilon(\boldsymbol{E}) + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\\ &\quad + \boldsymbol{\beta}^T\mathbb{E}_\epsilon(\boldsymbol{E})^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\mathbb{E}_\epsilon(\boldsymbol{E})\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbb{E}_\epsilon(\boldsymbol{E}^T\boldsymbol{E})\boldsymbol{\beta})\\ &= \frac{1}{2}(\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{\beta}\boldsymbol{X} - 2\boldsymbol{Y}^T\boldsymbol{\beta}\boldsymbol{0} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\\ &\quad + \boldsymbol{\beta}^T\boldsymbol{0}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{0}\boldsymbol{\beta} + \sigma^2\boldsymbol{\beta}^T\boldsymbol{I}\boldsymbol{\beta})\\ &= \frac{1}{2}\left(\boldsymbol{Y}-\boldsymbol{\beta}\boldsymbol{X}\right)^T(\boldsymbol{Y}-\boldsymbol{\beta}\boldsymbol{X}) + \frac{\sigma^2}{2}\|\boldsymbol{\beta}\|^2\end{aligned}$$

   que es equivalente a Ridge con $\lambda = \frac{\sigma^2}{2}$.

3. Con predictores ortogonales y asumiendo que la matrix $\boldsymbol{X}$ está centrada tenemos que

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{\Sigma} &= \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)\\ \boldsymbol{\Sigma}^{-1} &= \mathrm{diag}(\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_d^{-2})\,,\end{aligned}$$

   La solución de mínimos cuadrados es

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

   Si denotamos por $\boldsymbol{X}_{(j)}$ al vector correspondiente a la j-ésima columna de $\boldsymbol{X}$ tenemos que el j-ésimo coeficiente de $\hat{\boldsymbol{\beta}}$ es

$$\hat{\boldsymbol{\beta}}_j = \frac{1}{\sigma_j^2}\boldsymbol{X}_{(j)}^T\boldsymbol{Y}\,,$$

y su z-score viene dado por

$$z_j = \frac{\hat{\boldsymbol{\beta}}_j}{\hat{\sigma}/\sigma_j} = \frac{\hat{\boldsymbol{\beta}}_j \sigma_j}{\hat{\sigma}} \,,$$

Si dos coeficientes $\hat{\boldsymbol{\beta}}_j$ y $\hat{\boldsymbol{\beta}}_i$ tienen el mismo peso

$$\frac{1}{\sigma_j^2} \boldsymbol{X}_{(j)}^T \boldsymbol{Y} = \frac{1}{\sigma_i^2} \boldsymbol{X}_{(i)}^T \boldsymbol{Y}$$

De estos dos coeficientes, aquel con mayor z-score es aquél con mayor σ_j (o, equivalentemente, mayor σ_j^2). Ahora, la solución de Ridge es

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

cuyo j-ésimo coeficiente es

$$\tilde{\boldsymbol{\beta}}_j = \frac{1}{\sigma_j^2 + \lambda} \boldsymbol{X}_{(j)}^T \boldsymbol{Y}$$

Tenemos que demostrar que si $\sigma_j^2 > \sigma_i^2$,

$$\frac{1}{\sigma_j^2 + \lambda} \boldsymbol{X}_{(j)}^T \boldsymbol{Y} > \frac{1}{\sigma_i^2 + \lambda} \boldsymbol{X}_{(i)}^T \boldsymbol{Y}$$

o equivalentemente, que

$$1 = \frac{\frac{1}{\sigma_j^2} \boldsymbol{X}_{(j)}^T \boldsymbol{Y}}{\frac{1}{\sigma_i^2} \boldsymbol{X}_{(i)}^T \boldsymbol{Y}} > \frac{\sigma_i^2}{\sigma_j^2} \frac{(\sigma_j^2 + \lambda)}{(\sigma_i^2 + \lambda)}$$

En efecto, para λ estrictamente positivo,

$$\begin{aligned} \sigma_j^2 > \sigma_i^2 \Rightarrow \lambda\sigma_j^2 > \lambda\sigma_i^2 &\Rightarrow \sigma_j^2\sigma_i^2 + \lambda\sigma_j^2 > \sigma_j^2\sigma_i^2 + \lambda\sigma_i^2 \\ &\Leftrightarrow \sigma_j^2(\sigma_i^2 + \lambda) > \sigma_i^2(\sigma_j^2 + \lambda) \\ &\Leftrightarrow \frac{\sigma_i^2(\sigma_j^2 + \lambda)}{\sigma_j^2(\sigma_i^2 + \lambda)} < 1 \end{aligned}$$