Lecture_notes5

March 1, 2021

1 MATH310 - Lecture_notes5

1.1 Some particular bi-objective problems (see §15.5.2 in VMLS)

,

We focus on solving problems of the following type:

Find $\mathbf{x} \in \mathbb{R}^n$ minimizing the bi-objective function

$$J(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x} - \mathbf{x}^{des}\|^2,$$

where the $m \times n$ coefficient matrix A is "wide" (meaning that $n > m$ i.e. we have more unknowns than equations in the system $A\mathbf{x} = \mathbf{b}$) and the magnitude of $\lambda > 0$ indicate the strength in our desire for the solution $\mathbf{x}$ to be close to some (desired) $\mathbf{x}^{des} \in \mathbb{R}^n$.

With $A_1 = A$, $\mathbf{b}_1 = \mathbf{b}$, $A_2 = I_n$, $\mathbf{b}_2 = \mathbf{x}^{des}$, $\lambda_1 = 1$ and $\lambda_2 = \lambda$ the above bi-objective function can be expressed as

$$J(\mathbf{x}) = \lambda_1\|A_1\mathbf{x} - \mathbf{b}_1\|^2 + \lambda_2\|A_2\mathbf{x} - \mathbf{b}_2\|^2,$$

i.e. a *weigthed sum objective* of a bi-objective least squares problem, see §15.1 in VMLS.

1.2 An OLS-formulation of the above problem-type

,

Note that the above objective function $J(\mathbf{x})$ corresponds to the ordinary least squares (OLS) formulation for solving the system

$$\begin{bmatrix} A \\ \sqrt{\lambda} I_n \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{b} \\ \sqrt{\lambda}\mathbf{x}^{des} \end{bmatrix}$$

with the corresponding normal equations

$$(A^t A + \lambda I_n)\mathbf{x} = A^t\mathbf{b} + \lambda\mathbf{x}^{des}.$$

The least squares solution of this system is

$$
\begin{aligned}
\hat{\mathbf{x}} &= (A^tA + \lambda I_n)^{-1}(A^t\mathbf{b} + \lambda\mathbf{x}^{des}) \\
&= (A^tA + \lambda I_n)^{-1}(A^t\mathbf{b} + (\lambda I_n + A^tA)\mathbf{x}^{des} - (A^tA)\mathbf{x}^{des}) \\
&= (A^tA + \lambda I_n)^{-1}A^t(\mathbf{b} - A\mathbf{x}^{des}) + \mathbf{x}^{des}.
\end{aligned}
$$

Note that the inverted matrix $(A^tA + \lambda I_n)^{-1} \in \mathbb{R}^{n\times n}$.

1.3 The "kernel trick" for faster solution of the above problem type

,

Note that

$$(A^tA + \lambda I_n)A^t = A^t(AA^t + \lambda I_m),$$

where both $(A^tA + \lambda I_n)$ and $(AA^t + \lambda I_m)$ are invertible matrices for $\lambda > 0$.

Multiplication of the above equation from the left by $(A^tA + \lambda I_n)^{-1}$ and from the right by $(AA^t + \lambda I_m)^{-1}$ yields the identity

$$A^t(AA^t + \lambda I_m)^{-1} = (A^tA + \lambda I_n)^{-1}A^t.$$

Therefore the OLS solution of $\begin{bmatrix} A \\ \sqrt{\lambda} I_n \end{bmatrix}\mathbf{x} = \begin{bmatrix} \mathbf{b} \\ \sqrt{\lambda}\mathbf{x}^{des} \end{bmatrix}$ can also be expressed as

$$\hat{\mathbf{x}} = A^t(AA^t + \lambda I_m)^{-1}(\mathbf{b} - A\mathbf{x}^{des}) + \mathbf{x}^{des}.$$

Note that here the inverted matrix $(AA^t + \lambda I_m)^{-1} \in \mathbb{R}^{m\times m}$, which is a smaller problem for wide matrices $(n > m)$.

If $QR = \bar{A} = \begin{bmatrix} A^t \\ \sqrt{\lambda} I_m \end{bmatrix}$ is *the qr-decomposition* of the stacked $(n+m)\times m$ matrix $\bar{A}$. Then

$$(AA^t + \lambda I_m) = \bar{A}^t\bar{A} = R^tQ^tQR = R^tR,$$

and the OLS solution becomes

$$\hat{\mathbf{x}} = A^t(R)^{-1}(R^t)^{-1}(\mathbf{b} - A\mathbf{x}^{des}) + \mathbf{x}^{des}.$$

1.4 Tikhonov regularization (Ridge regression) modelling

,

Let's convert to "statistics notation" where X denotes a mean centered data matrix of size $m \times n$ where typically $n > m$ (we have more variables/unknowns than samples), $\mathbf{y}_0$ is the corresponding mean centered response and $\lambda > 0$. Then, if the "desired" solution of the above problem type is set to $\mathbf{0}$, our minimization problem is about finding $\beta \in \mathbb{R}^n$ minimizing the objective

$$J(\beta) = \|X\beta - \mathbf{y}_0\|^2 + \lambda\|\beta\|^2.$$

The corresponding OLS-problem is

$$\begin{bmatrix} X \\ \sqrt{\lambda} I_n \end{bmatrix} \beta = \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{y}_0 = \mathbf{y} - \bar{y}$ ($\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$) is the mean centered version of $\mathbf{y}$.

This type of OLS-problem is often called Tikhonov regularization (TR) or Ridge regression (RR), see §15.3.1 and §15.4 in VMLS.

According to the above derivations, the least squares solution of such problems is given by

$$\beta_\lambda = X^t(XX^t + \lambda I_m)^{-1}\mathbf{y}_0.$$

Analogously to PCR (see last weeks notes) we predict the response value $\hat{y}$ for a new datapoint (sample) $\mathbf{x}^t \in \mathbb{R}^n$ based on the λ-regularized **RR-model** by including a constant term $\beta_{0,k}$ to calculate

$$\hat{y} = \beta_{0,\lambda} + \mathbf{x}^t\beta_\lambda.$$

Here $\beta_{0,\lambda} = \bar{y} - \bar{\mathbf{x}}^t\beta_\lambda$ where $\bar{\mathbf{x}}^t$ is the (row) vector of column means used for centering of the data matrix X.

Note that for the particular choice $\mathbf{x} = \bar{\mathbf{x}}$ we obtain the prediction

$$\hat{y} = \beta_{0,\lambda} + \bar{\mathbf{x}}^t\beta_\lambda = \bar{y} - \bar{\mathbf{x}}^t\beta_\lambda + \bar{\mathbf{x}}^t\beta_\lambda = \bar{y},$$

i.e. from the mean of the observed X-data we predict the mean of the observed $\mathbf{y}$-data, just as we did for the PCR-models.

1.5 Model validation and -selection

,

Question: How do we select the number of principal components (k) in PCR and the regularization parameter value (λ) i RR to obtain models with good predictions?

Answer: We can do 10-fold cross validation or leave-one-out cross validation (recall §13.2 in VMLS) for the various candidate models, compare the RMS-values for the predictions to choose a model with seemingly low prediction error…