TU München Summersemester 2019

Prof. Dr. Dr. Fabian Theis & Dr. Tingying Peng

MA4002: Statistical Learning

Sheet 3: bias-variance tradeoff / ridge regression

Exercise 1: Assuming a linear model

$$P(y|x,\beta,\sigma^2) = \mathcal{N}(y|x^T\beta,\sigma^2),$$

show that the variance (covariance) of the least-squares estimator $\hat{\beta}$ of β is

$$\mathrm{var}\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.$$

Exercise 2: Analyze the relation of prostate-specific antigen and clinical measures among men who were about to have their prostates removed (`prostate.data`) using ordinary least-squares and Ridge regression. The meaning of features and the predicted variable `lpsa` are as follows

- lcavol : log(cancer volume)
- lweight : log(prostate weight)
- age : age
- lbph : log(benign prostatic hyperplasia)
- svi : seminal vesicle invasion
- lcp : log(capsular penetration)
- gleason : Gleason score
- pgg45 : percent of Gleason scores 4 or 5
- lpsa : log(prostate-specific antigen)

`prostate.data` has a column marked `train` (boolean variable) that allows you to get train and test set.

(a) Scale the data to zero mean and unit variance using the `preprocessing.scale` function.

(b) Fit a ordinary linear model using the `linear_model.LinearRegression` function. Compute the l2-norm of error on the test set. It should be noted that we also need to calculate intercept for the model (setting `fit_intercept_=True`).

(c) Fit a Ridge regression using `linear_model.Ridge` function argument λ= `np.logspace(-10, 3, 1000)`. Plot the model parameters $\hat{\beta}$ versus λ.

(d) The scale of λ is not very intuitive. Therefore plot $\hat{\beta}$ versus the number of effective degrees of freedom

$$\mathrm{df}(\lambda) = \mathrm{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T].$$

where $\mathbf{I}$ is the identity matrix.

(e) Use *cross validation* to determine the optimal value for λ (using `linear_model.RidgeCV` function). And compare the performance of the ordinary linear regression model and the Ridge regression model with the optimal λ on the test set.

Exercise 3: (optional) Suppose we have an i.i.d. training set $S = \{(x_i, y_i)\}_{i=1}^N \subset X \times Y$, where $x_i \sim p(x)$, $y_i = f(x_i) + \varepsilon_i$ and $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ is a noise vector, $\varepsilon_i = N(0, \sigma^2)$. We construct an estimator for f as follows:

$$f_S(\cdot) = \sum_{i=1}^N \omega_i(\cdot, X) y_i, \tag{1}$$

where the weights $\omega_i(\cdot, X)$ only depend on X.

- Show that both linear regression and $k-$nearest-neighbor regression belong to this type. Describe explicitly the weights $\omega_i(\cdot, X)$ in each of these cases.
- Show that the *conditional mean-squared error* can be evaluated as

  $$E_{Y|X}(f(x_0) - f_S(x_0))^2 = \mathtt{Bias}^2(f_S(x_0)) + \mathtt{Var}(f_S(x_0)).$$

  Evaluate the conditional mean-squared error in terms of the weights $\omega_i(\cdot, X)$ and the noise distribution, ε_i.
- Decompose the mean-squared error $E_{Y,X}(f(x_0) - f_S(x_0))^2$ into a squared bias and a variance component.