# 36700 – Probability and Mathematical Statistics

*Spring 2019*

## Homework 9

*Due Friday, April 26th at 12:40 PM*

- All homework assignments shall be uploaded using Gradescope through the Canvas portal). Late submissions are not allowed.

1. In nonparametric regression, let $\hat{r}(x)$ be an estimate of $r(x) = \mathbb{E}(Y|X = x)$, with bias $b(x) = \mathbb{E}\hat{r}(x) - r(x)$ and variance $v(x) = \mathrm{Var}(\hat{r}(x))$. Define $R(\hat{r}, r) = \int \mathbb{E}[\hat{r}(x) - r(x)]^2 dx$. Prove that $R(\hat{r}, r) = \int b^2(x)dx + \int v(x)dx$.
   Indeed, this decomposition also holds for any function estimation, such as density estimation.

2. Consider linear regression model
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
   and assume that $\mathbf{X}$ is orthonormal in the sense that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. Assume also that $\boldsymbol{\epsilon}$ has mean zero and covariance $\sigma^2\mathbf{I}_n$. Let $\hat{\boldsymbol{\beta}}_{\text{ridge}}(\alpha)$ be the ridge regression estimate with penalty parameter $\alpha$. Derive the bias $b = \mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{ridge}}(\alpha) - \boldsymbol{\beta}]$, and variance $v = \mathrm{Var}[\hat{\boldsymbol{\beta}}_{\text{ridge}}(\alpha)]$. Note that here $b$ is a $p \times 1$ vector and $v$ is a $p \times p$ covariance matrix.

3. In the previous problem, assume that
$$\mathbf{X}^T\mathbf{X} = \mathrm{diag}(\tau_1, ..., \tau_p)$$
   (i.e., a diagonal matrix with diagonal entries being $\tau_1, ..., \tau_p$). Derive the bias and variance of the ridge regression estimate, and comment on the bias-variance trade-off when some of the $\tau_j$'s are close to zero.

4. Suppose in the kernel non-parametric regression problem we want to estimate $r(X_i)$ for each $i = 1, ..., n$. Then the kernel regression estimate would be
$$\hat{y}_i = \hat{r}(X_i) = \frac{\sum_{j=1}^n K\left(\frac{X_j - X_i}{h}\right) Y_j}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}$$
   which can be viewed as a linear smoother in the matrix form
$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{Y}$$
   where
$$\mathbf{H}_{ij} = \frac{K\left(\frac{X_j - X_i}{h}\right)}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}.$$
   Derive the GCV formula (as a function of the data, kernel $K(\cdot)$, bandwidth $h$).

5. Consider the same data set as in Q6 of HW8.

   (a) Fit a nonparametric regression estimate using five different values of $h : 1/8, 1/4, 1/2, 1, 2$. Plot the fitted curves using $(X_i, \hat{y}_i)$. You can pick either the Gaussian or or the box kernel.

   (b) Conduct model selection to pick the best bandwidth among the five values of $h$, using leave-out-out cross-validation and GCV, respectively. Report the values of $CV(h)$ and $GCV(h)$ for each $h$ (the latter will be based on your derivation in the previous exercise). Compare the results of CV and GCV.

   You are NOT required to submit your code for this problem.