Question 3.1 (Effective degrees of freedom): In class, we said that an estimator \( \hat{f} \) making predictions \( \hat{Y}_i = \hat{f}(x_i) \) from a linear model \( Y = X\beta + \varepsilon, \varepsilon \sim (0, \sigma^2 I) \), had effective degrees of freedom
\[
dof(\hat{f}) := \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(\hat{Y}_i, Y_i) = E[(Y - E[Y])^T \hat{Y}] .
\]
Treating \( X \) as a fixed design, compute the degrees of freedom for the following estimators:

(a) The principal components regression estimator using \( r \) components. Recall that this estimator is as follows (here, we shall assume \( X \) is standardized so that \( X^T 1 = 0 \) and \( n^{-1} \sum_{i=1}^{n} X_{ij}^2 = 1 \) for each \( j \in \{1, \ldots, d\} \): we compute the principal components of variation in \( X \), project the rows of \( X \) onto these, and perform ordinary least squares on the resulting low rank matrix. If \( X = U \Sigma V^T \) is the SVD of \( X \), then if \( U_r = [u_1 \cdots u_r], \Sigma_r = \text{diag}(s_1, \ldots, s_r), V_r = [v_1 \cdots v_r] \in \mathbb{R}^{d \times r} \), then this is equivalent to setting
\[
\hat{\beta}_{pcr}(r) = \arg\min_b \left\{ ||Y - XV_rV_r^T b||_2^2 \right\},
\]
or
\[
\hat{\gamma}_r = \arg\min_{\gamma \in \mathbb{R}^r} \left\{ ||Y - U_r \Sigma_r \gamma||_2^2 \right\}, \quad \hat{\beta}_{pcr}(r) = V_r \hat{\gamma}_r.
\]
(b) The ridge regression estimator, where
\[
\hat{\beta}_\lambda = \arg\min_b \left\{ ||X b - Y||_2^2 + \lambda \|b\|_2^2 \right\}.
\]
(In all cases, \( \hat{Y} = X \hat{\beta} \) for each estimator.)

Question 3.2 (Kernel machines): Kernel regression is essentially a generalization of locally weighted linear regression and allows us to apply regression ideas to any space; we will consider trying to predict responses \( y \in \mathbb{R} \) given \( x \in \mathcal{X} \), where \( \mathcal{X} \) is some space. One is given a kernel function \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \), which is a symmetric function with the property that the Gram matrix defined by
\[
G_{ij} = k(x_i, x_j), \quad G \in \mathbb{R}^{n \times n}
\]
is always positive semidefinite for any collection \( \{x_1, \ldots, x_n\} \) of vectors. Some typical choices of kernels include

i. the radial basis function or Gaussian kernel, which for a bandwidth \( \tau > 0 \) is
\[
k(x, z) = \exp\left(-\frac{||x - z||_2^2}{2\tau^2}\right),
\]
ii. polynomial kernels, which for a degree \( d \) are given for vectors \( x, z \in \mathbb{R}^p \) by
\[
k(x, z) = (1 + x^T z)^d,
\]
iii. the min-kernel for data \( x \in [0, 1] \), given by
\[
k(x, z) = \min\{x, z\}.
\]
One typically thinks of $k$ as measuring something like similarity between its inputs: large values of $k$ correspond to “close” vectors (though, as evidenced by the min kernel, this need not be precisely the case).

Consider the case of approximating $y = f(x) + \varepsilon$, $\varepsilon \sim (0, \sigma^2)$. Kernel regression approximates $f$ by

$$h(x) = \sum_{i=1}^{n} k(x, x_i) \alpha_i. \quad (3.1)$$

We define the norm of such a function $h$ by

$$\|h\|^2 := \alpha^T G \alpha$$

when $G$ is the Gram matrix.\(^1\) Kernel ridge regression then chooses $\hat{f}$ by minimizing

$$\sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda \|h\|^2 \quad (3.2)$$

over $h$ of the form (3.1), so $\hat{f}(x) = \sum_{i=1}^{n} k(x, x_i) \hat{\alpha}_i$ for some estimated $\hat{\alpha} \in \mathbb{R}^n$.

(a) Explain in words why kernel regression makes sense, and why the resulting function $\hat{f}$ is smooth as a function of its input $x$ whenever the kernel function $k$ is.

(b) Give an explicit formula for the coefficients $\hat{\alpha}$ minimizing the squared error (3.2). Is your solution unique even when $G$ is low rank?

(c) Let $\hat{y}_i = \hat{f}(x_i)$ for your estimator above. If $y = f(x) + \varepsilon$ for $\varepsilon \sim (0, \sigma^2)$ independent and mean zero, compute the effective degrees of freedom

$$\text{dof}(\hat{f}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(\hat{y}_i, y_i).$$

(d) For the Gaussian kernel in item i, give the limits of the effective degrees of freedom $\text{dof}(\hat{f})$ as (i) $\tau^2 \uparrow \infty$ and (ii) $\tau^2 \downarrow 0$, assuming the $x_i$ are all distinct. Interpret your results: if $\lambda \approx 0$ and $\tau \approx 0$, what are your predictions $\hat{y}_i$?

(e) Let $f = [f(x_1) \cdots f(x_n)]^T$ be the vector with entries $f(x_i)$, and let $\hat{y}$ be the vector of predicted values. Show that for small regularization $\lambda > 0$, the (in-sample) bias satisfies

$$f - \mathbb{E}[\hat{y}] = \lambda G^{-1} f + O(\lambda^2)$$

assuming that $G$ is invertible, where $O(\lambda^2)$ indicates an error term that tends to zero as fast as $\lambda^2$. You may use that if $A$ is a positive definite matrix, then $(A + \Delta)^{-1} = A^{-1} + \sum_{k=1}^{\infty} (-1)^k (A^{-1} \Delta) A^{-1}$ whenever $\|\Delta\|_{\text{op}} < \lambda_{\text{min}}(A)$.

(f) By question (e), for small $\lambda > 0$ the ridge regression estimator is (nearly) unbiased. Use this to argue that the expected residual sum of squares can give a natural variance estimate as follows. Let $H_\lambda = G(G + \lambda I)^{-1}$. Show that

$$\mathbb{E}[\text{RSS}] = \mathbb{E}[\|\hat{y} - y\|^2] = \sigma^2 \left( n - 2 \cdot \text{dof}(\hat{f}) + \text{tr}(H_\lambda^2) \right) + O(\lambda^2)$$

\(^1\)We attach a small appendix to this exercise set that is completely optional but may be of some interest to those of you who are analytically inclined; see Appendix A.
for small $\lambda$. Use this and part (e) to argue that

$$\hat{\sigma}^2 := \frac{1}{n - 2\text{dof}(\hat{f}) + \text{tr}(H_\lambda^2)} \|\hat{y} - y\|_2^2$$

is a “reasonable” estimate of $\sigma^2$ when $\lambda$ is small.

As a small aside, a careful calculation shows that $n - 2\text{dof}(\hat{f}) + \text{tr}(H_\lambda^2) = O(\lambda^2)$ when $\lambda$ is small, so sometimes a more involved approach is warranted. In this case, if we assume that $G$ is full rank—which will be the case, for example, when one uses the Gaussian kernel $k(x, z) = \exp\left(-\frac{1}{2\tau^2} \|x - z\|_2^2\right)$—we can write $f = G\alpha^*$ for some $\alpha^* \in \mathbb{R}^n$. Then for small $\lambda$ we have

$$\text{E}[\text{RSS}] = \text{E}[\|\hat{y} - y\|_2^2] = \sigma^2 \left(n - 2 \cdot \text{dof}(\hat{f}) + \text{tr}(H_\lambda^2)\right) + \lambda^2 \|\alpha^*\|_2^2 + O(\lambda^3).$$

Thus, if we can estimate $\|\alpha^*\|_2^2$, we can get some further corrections; a reasonable heuristic here is to substitute the $\|\cdot\|_2$-norm of $G^{-1}\hat{y}$ for $\alpha^*$, then use the estimate

$$\tilde{\sigma}^2 := \frac{1}{n - 2\text{dof}(\hat{f}) + \text{tr}(H_\lambda^2)} \left[\|\hat{y} - y\|_2^2 - \lambda^2 \|G^{-1}\hat{y}\|_2^2\right].$$

(Sometimes this may be non-positive, and so further corrections may be necessary.)

**Question 3.3** (Implementing and experimenting with kernel ridge regression): In this question, you will use the data available in *lprostate.dat* from the paper [3] to investigate and implement various versions of kernel ridge regression. In this dataset, the goal is to predict the log prostate specific antigen *lpsa* from other prostate measurements.

(a) Implement a function `predictKRR` that takes as input five quantities:

- $X$, a $d \times n$ data matrix representing the training data with rows $x_i^T$
- $Z$, a $d \times m$ matrix of $m$ data points $z_1, \ldots, z_m$ on which to make new predictions
- $\text{alpha}$, an $n$ vector representing the vector $\hat{\alpha}$ fit in Part (b) of Question 3.2
- $\tau$, the bandwidth parameter $\tau$ for the Gaussian (RBF) kernel $k(x, z) = \exp\left(-\frac{1}{2\tau^2} \|x - z\|_2^2\right)$
- $\text{offset}$, an offset parameter $b \in \mathbb{R}$

The function should then output a vector of predictions $\hat{y}(z_j) = b + \sum_{i=1}^n k(x_i, z_j)\alpha_i$, where $k$ is the Gaussian kernel with the given bandwidth, $k(x, z) = \exp\left(-\frac{\|x - z\|_2^2}{2\tau^2}\right)$

(b) Implement a method `fitKRR` that takes as input

- $X$, a $d \times n$ data matrix representing the training data with rows $x_i^T$
- $y$, an $n$ vector of responses
- $\text{lambda}$, a positive scalar representing regularization
- $\tau$, the bandwidth parameter $\tau$

then returns the pair $(\text{alpha}, \text{yMean})$, where $\text{alpha}$ is the solution $\hat{\alpha}$ to the kernel ridge regression problem (3.2) with the Gaussian kernel where we replace $y$ with $y - \bar{y}$, the centered $y$ (here $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$), and $\text{yMean}$ is the mean response.
(c) Perform kernel ridge regression on the data in `lprostate.dat`, using `lpsa` as the response and use `lcavol` as the only covariate `x`. You should standardize the data in `x` first so that the (sample) variance of `x` is 1. Fix $\tau = .1$ and vary $\lambda \in \{.01, .5, 5\}$. Plot the resulting fitted predictions $\hat{y}$ superimposed on a scatter plot of `lcavol` against `lpsa`. Describe what you see in words.

(d) Repeat an identical experiment to that above, but use $\tau = .5$.

(e) Now we use all the covariates to predict `lpsa`. Perform kernel ridge regression using `lprostate.dat`. Standardize the data (excepting `lpsa`), and for $\tau = .5$, fit a kernel ridge regression for $\lambda \in \{.01, .5, 5\}$. Plot the resulting fitted predictions $\hat{y}$ superimposed on a scatter plot of `lcavol` against `lpsa`. Describe what you see in words.

(f) Find the estimate $\hat{\sigma}^2$ defined in part (f) of question 3.2 using $\tau = .5$ and $\lambda = .1$, using all covariates (standardized) as in the preceding part. (Recall as in class that estimating $\hat{\sigma}$ with a “more powerful” model is often a good way to estimate the smallest irreducible error.) Now, perform kernel ridge regression as in part (d) using only `lcavol` as the covariate with $\tau = .5$ and $\lambda = .5$. Estimate the standard error of the fit for each predicted value $\hat{y}_i$, and superimpose your predictions $\hat{y}$ with 2× standard error bands on a scatter of `lcavol` and `lpsa`.

### A Reproducing kernel Hilbert spaces

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite if for any $n \in \mathbb{N}$ and collection of points $x_1, \ldots, x_n$, the Gram matrix $G \in \mathbb{R}^{n \times n}$ with entries $G_{ij} = k(x_i, x_j)$ is positive semidefinite. Any such $k$ gives rise to a Hilbert space $\mathcal{H}$ of functions from $\mathcal{X} \to \mathbb{R}$ for which $k$ is the reproducing kernel (sometimes called the representer of evaluation), meaning that for any $h \in \mathcal{H}$, we have

$$\langle h, k(x, \cdot) \rangle = h(x),$$

as follows. For any finite $n, m$, and $x_i, z_i \in \mathcal{X}$, we define the inner product between $h_0(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$ and $h_1(x) = \sum_{i=1}^m \beta_i k(z_i, x)$ by

$$\langle h_0, h_1 \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, z_j).$$

It is evident that if $h(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$, then $\langle h, k(x, \cdot) \rangle = h(x)$. By taking $\mathcal{H}$ to be the completion of this space with the given inner product, we then have a complete inner product space with norm $\|h\|^2 = \langle h, h \rangle$. If $h$ has the form $h(x) = \sum_{i=1}^n k(x, x_i) \alpha_i$ and $G = [k(x_i, x_j)]_{i,j} \in \mathbb{R}^{n \times n}$ is the Gram matrix, then the norm evidently satisfies

$$\|h\|^2 = \langle h, h \rangle = \alpha^T G \alpha.$$

Note that our construction of $\mathcal{H}$ as the completion via the reproducing kernel $k$ shows that for any such positive semidefinite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, z) = \langle \phi(x), \phi(z) \rangle,$$

so that $k$ can represent a (potentially) infinite dimensional inner product. Additionally, for any feature mapping $\phi(x)$, the function defined by $k_{\phi}(x, z) = \langle \phi(x), \phi(z) \rangle$ evidently is positive semidefinite and gives rise to a Hilbert space in the same way as above.
The representer theorem

We can then frame the optimization problem (3.2) in a more general way: consider the problem
to minimize \( h \in H \)

\[
\sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda \| h \|^2,
\]

(A.2)

where we just let \( h \in H \). Now, we can write any \( h \in H \) as \( h(x_i) = \sum_{i=1}^{n} \alpha_i k(x_i, x_i) + g(x_i) \) and \( g, k(x_i, \cdot) = 0 \) for each \( i \) by orthogonalizing. In this case,

\[
\| h \|^2 = \alpha^T G \alpha + \| g \|^2,
\]

while \( h(x_i) = \sum_{j=1}^{n} \alpha_i k(x_j, x_i) + (k(x_i, \cdot), g) = \sum_{j=1}^{n} \alpha_i k(x_j, x_i) \). Thus including such a non-zero \( g \) can only increase \( \| h \|^2 \) without modifying \( h(x_i) \), and the minimizer of problem (A.2) is necessarily of the form

\[
h(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).
\]

This result is called the representer theorem. There are whole books devoted to the topic; see, for example, [4, 2, 1]. A major theme in this line of work is to develop efficient and effective kernel functions \( k \) for problems of interest.

Minimizing squared error with a feature mapping

The representer theorem, coupled with the feature mapping (A.1), give an alternative interpretation of kernel ridge regression (A.2): we can instead consider the problem

\[
\min_{\beta} \sum_{i=1}^{n} (y_i - \langle \phi(x_i), \beta \rangle)^2 + \lambda \| \beta \|^2
\]

(A.3)

(where \( \beta \) is in the same space as \( \phi(x_i) \)). Then by the representer theorem, for any \( \hat{\beta} \) minimizing this objective there is some \( \alpha \in \mathbb{R}^n \) such that

\[
\langle \hat{\beta}, \phi(x) \rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x_i),
\]

and so the problem (A.3) is equivalent to kernel ridge regression (A.2), and so also to problem (3.2). As a side benefit to this perspective, we see how to include an intercept term in problem (3.2) with kernel \( k \): let \( \phi \) be the feature mapping associated to the kernel, so \( k(x, z) = \langle \phi(x), \phi(z) \rangle \). Then evidently, replacing \( \phi \) with

\[
\phi_{\text{int}}(x) = \begin{bmatrix} 1 \\ \phi(x) \end{bmatrix}
\]

gives rise to the kernel function \( k_{\text{int}}(x, z) = 1 + \langle \phi(x), \phi(z) \rangle = 1 + k(x, z) \). Then instead of solving problem (3.2), we solve

\[
\min_{\alpha \in \mathbb{R}^n} \| y - \alpha^T (11^T + G) \alpha \|^2 + \lambda \alpha^T G \alpha,
\]

where \( G = [k(x_i, x_j)]_{i,j} \) is the usual Gram matrix for \( k \).
References


