A woodchuck is peeking out of a hole in a snowy landscape. The woodchuck's head and shoulders are visible, and it is looking directly at the camera. The background is a snowy field with some dry grass and a few brown leaves scattered around.

S&DS 365 / 665
Intermediate Machine Learning

Smoothing and Kernels

Wednesday, February 2

Please note

- Changes to calendar <http://interml.ydata123.org>
- Readings from “Probabilistic Machine Learning: An Introduction”
- <https://probml.github.io/pml-book/book1.html>

Topics for today

- Recap: Smoothing methods
- Demo of smoothing with various kernels
- Mercer kernels

Nonparametric Regression

Given $(X_1, Y_1), \dots, (X_n, Y_n)$ predict Y from X .

Assume only that $Y_i = m(X_i) + \epsilon_i$ where $m(x)$ is a smooth function of x .

The most popular methods are *kernel methods*. However, there are two types of kernels:

- 1 Smoothing kernels
- 2 Mercer kernels

Smoothing kernels involve local averaging.
Mercer kernels involve regularization.

Smoothing Kernels

- Smoothing kernel estimator:

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n Y_i K_h(X_i, x)}{\sum_{i=1}^n K_h(X_i, x)}$$

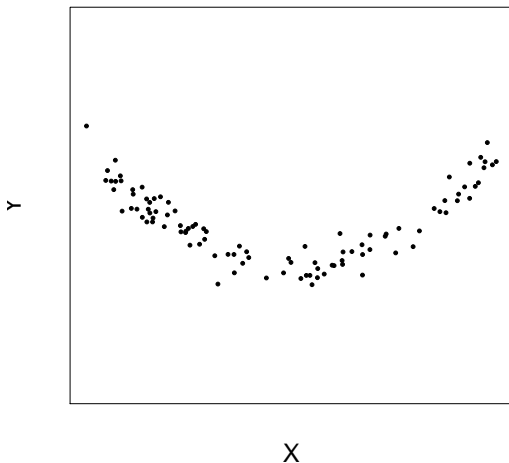
where $K_h(x, z)$ is a *kernel* such as

$$K_h(x, z) = \exp\left(-\frac{\|x - z\|^2}{2h^2}\right)$$

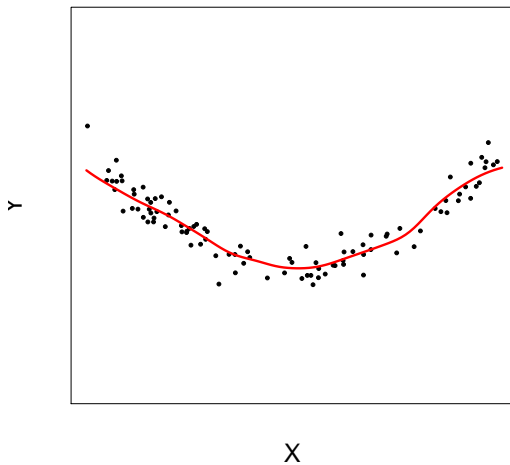
and $h > 0$ is called the *bandwidth*.

- $\hat{m}_h(x)$ is just a local average of the Y_i 's near x .
- The bandwidth h controls the bias-variance tradeoff:
Small h = large variance while *large h = large bias*.

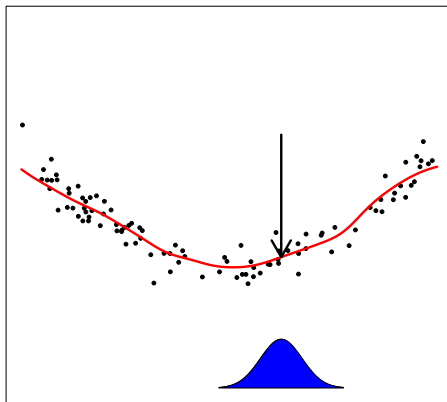
Example: Some Data – Plot of Y_i versus X_i



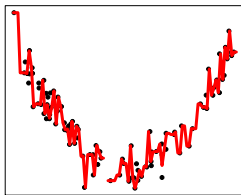
Example: $\hat{m}(x)$



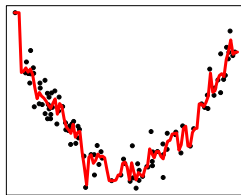
$\hat{m}(x)$ is a local average



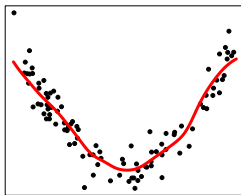
Effect of the bandwidth h



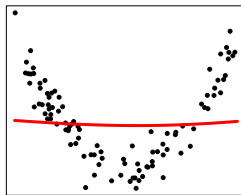
very small bandwidth



small bandwidth



medium bandwidth



large bandwidth

Smoothing Kernels

$$\text{Risk} = \mathbb{E}(Y - \hat{m}_h(X))^2 = \text{bias}^2 + \text{variance} + \sigma^2.$$

$$\text{bias}^2 \approx h^4,$$

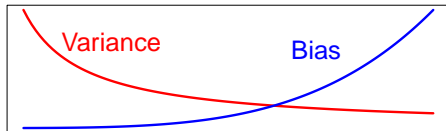
$$\text{variance} \approx \frac{1}{nh^p} \quad \text{where } p = \text{dimension of } X.$$

$\sigma^2 = \mathbb{E}(Y - m(X))^2$ is the unavoidable prediction error.

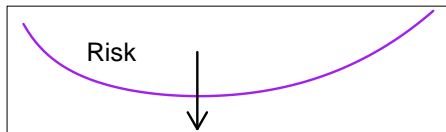
small h: low bias, high variance (undersmoothing)

large h: high bias, low variance (oversmoothing)

Risk Versus Bandwidth



h



optimal h

Estimating the Risk: Cross-Validation

To choose h we need to estimate the risk $R(h)$. We can estimate the risk by using *cross-validation*.

- 1 Omit (X_i, Y_i) to get $\hat{m}_{h,(i)}$, then predict: $\hat{Y}_{(i)} = \hat{m}_{h,(i)}(X_i)$.
- 2 Repeat this for all observations.
- 3 The cross-validation estimate of risk is:

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_{(i)})^2.$$

Shortcut formula: Whenever $\hat{Y} = LY$ we can use the shortcut

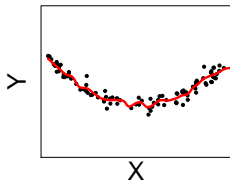
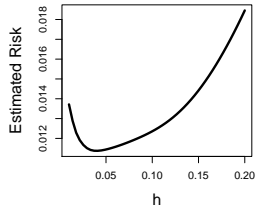
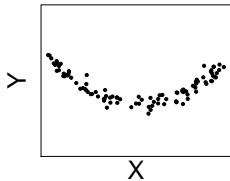
$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - L_{ii}} \right)^2.$$

In this case $L_{ij} = K_h(X_i, X_i) / \sum_t K_h(X_i, X_t)$.

Summary so far

- 1 Compute \hat{m}_h for each h .
- 2 Estimate the risk $\hat{R}(h)$.
- 3 Choose bandwidth \hat{h} to minimize $\hat{R}(h)$.
- 4 Let $\hat{m}(x) = \hat{m}_{\hat{h}}(x)$.

Example



Let's revisit the notebook

Another Approach: Mercer Kernels

Instead of using local smoothing, we can optimize the fit to the data subject to regularization (penalization). Choose \hat{m} to minimize

$$\sum_i (Y_i - \hat{m}(X_i))^2 + \lambda \text{penalty}(\hat{m})$$

where $\text{penalty}(\hat{m})$ is a *roughness penalty*.

λ is a parameter that controls the amount of smoothing.

How do we construct a penalty that measures roughness? One approach is: *Mercer Kernels* and *RKHS = Reproducing Kernel Hilbert Spaces*.

What is a Mercer Kernel?

A *Mercer Kernel* $K(x, x')$ is symmetric and positive definite:

$$\int \int f(x)f(x')K(x, x') dx dx' \geq 0 \quad \text{for all } f.$$

Example: $K(x, x') = e^{-\|x-x'\|^2/2}$.

Think of $K(x, x')$ as the *similarity* between x and x' . We will create a set of *basis functions* based on K .

Fix z and think of $K(z, x)$ as a function of x . That is,

$$K(z, x) = K_z(x)$$

is a function of the second argument, with the first argument fixed.

Mercer Kernels

Let

$$\mathcal{F} = \left\{ f(\cdot) = \sum_{j=1}^k \beta_j K(z_j, \cdot) \right\}$$

Define a norm: $\|f\|_K = \sum_j \sum_k \beta_j \beta_k K(z_j, z_k)$. $\|f\|_K$ *small means f smooth*.

If $f = \sum_r \alpha_r K(z_r, \cdot)$, $g = \sum_s \beta_s K(w_s, \cdot)$, the inner product is

$$\langle f, g \rangle_K = \sum_r \sum_s \alpha_r \beta_s K(z_r, w_s).$$

\mathcal{F} is a **reproducing kernel Hilbert space (RKHS)** because

$$\langle f, K(x, \cdot) \rangle = f(x)$$

Nonparametric Regression: Mercer Kernels

Representer Theorem

Let \hat{m} minimize

$$J(m) = \sum_{i=1}^n (Y_i - m(X_i))^2 + \lambda \|m\|_K^2.$$

Then

$$\hat{m}(x) = \sum_{i=1}^n \alpha_i K(X_i, x)$$

for some $\alpha_1, \dots, \alpha_n$.

So, we only need to find the coefficients

$$\alpha = (\alpha_1, \dots, \alpha_n).$$

Nonparametric Regression: Mercer Kernels

Plug $\hat{m}(x) = \sum_{i=1}^n \alpha_i K(X_i, x)$ into J :

$$J(\alpha) = \|Y - \mathbb{K}\alpha\|^2 + \lambda \alpha^T \mathbb{K}\alpha$$

where $\mathbb{K}_{jk} = K(X_j, X_k)$

Now we find α to minimize J . We get: $\hat{\alpha} = (\mathbb{K} + \lambda I)^{-1} Y$ and $\hat{m}(x) = \sum_i \hat{\alpha}_i K(X_i, x)$.

The estimator depends on the amount of regularization λ . Again, there is a bias-variance tradeoff. We choose λ by cross-validation. This is like the bandwidth in smoothing kernel regression.

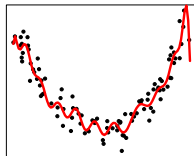
Smoothing Kernels *Versus* Mercer Kernels

Smoothing kernels: the bandwidth h controls the amount of smoothing.

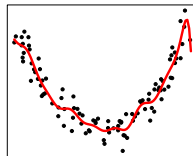
Mercer kernels: norm $\|f\|_K$ controls the amount of smoothing.

In practice these two methods give answers that are very similar.

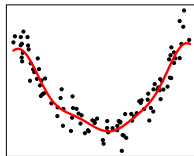
Mercer Kernels: Examples



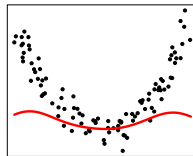
very small lambda



small lambda



medium lambda



large lambda

Multiple Regression

Both methods extend easily to the case where X has dimension $p > 1$. For example, just use

$$K(x, x') = e^{-\|x-x'\|^2/2}.$$

However, this is hard to interpret and is subject to the **curse of dimensionality**. This means that the *statistical performance* and the *computational complexity* degrade as dimension p increases.

An alternative is to use something less nonparametric such as **additive model** where we restrict $m(x_1, \dots, x_p)$ to be of the form:

$$m(x_1, \dots, x_p) = \beta_0 + \sum_j m_j(x_j).$$

Additive Models

Model: $m(x) = \beta_0 + \sum_{j=1}^p m_j(x_j)$.

We can take $\hat{\beta}_0 = \bar{Y}$ and we will ignore β_0 from now on.

We want to minimize

$$\sum_{i=1}^n \left(Y_i - (m_1(X_{i1}) + \cdots + m_p(X_{ip})) \right)^2$$

subject to m_j smooth.

Additive models: Backfitting Algorithm

Input: Data (X_i, Y_i)

Iterate until convergence:

For each $j = 1, \dots, p$:

Compute residual: $R_j = Y - \sum_{k \neq j} \hat{m}_k(X_k)$

Smooth $\hat{m}_j = S_j R_j$

Output: Estimator $\hat{m}(X_i) = \sum_j \hat{m}_j(X_{ij})$.

Here, $S_j R$ is any 1-dimensional nonparametric regression smoother

But what if p is large?

Sparse Additive Models

Additive Model: $Y_i = \sum_{j=1}^p m_j(X_{ij}) + \varepsilon_i, \quad i = 1, \dots, n$

High dimensional: $n \ll p$, with most $m_j = 0$.

Optimization:

$$\begin{aligned} & \text{minimize} && \mathbb{E} \left(Y - \sum_j m_j(X_j) \right)^2 \\ & \text{subject to} && \sum_{j=1}^p \sqrt{\mathbb{E}(m_j^2)} \leq L_n \\ & && \mathbb{E}(m_j) = 0 \end{aligned}$$

This generalizes the lasso!

Sparse Backfitting Algorithm

Input: Data (X_i, Y_i) , regularization parameter λ .

Iterate until convergence:

For each $j = 1, \dots, p$:

Compute residual: $R_j = Y - \sum_{k \neq j} \hat{m}_k(X_k)$

Smooth $\hat{m}_j = \mathcal{S}_j R_j$

Estimate norm: $s_j = \sqrt{\mathbb{E}(\hat{m}_j^2)}$

Soft-threshold: $\hat{m}_j \leftarrow \left[1 - \frac{\lambda}{\hat{s}_j} \right]_+ \hat{m}_j$

Output: Estimator $\hat{m}(X_i) = \sum_j \hat{m}_j(X_{ij})$.

This generalizes coordinate descent algorithm from last time.

Example: Boston Housing Data

Predict house value Y from 10 covariates.

We added 20 irrelevant (random) covariates to test the method.

Y = house value; $n = 506$, $p = 30$.

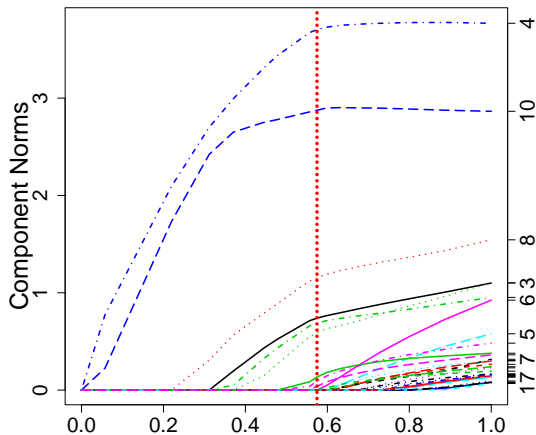
$$Y = \beta_0 + m_1(\text{crime}) + m_2(\text{tax}) + \cdots + \cdots m_{30}(X_{30}) + \epsilon.$$

Note that $m_{11} = \cdots = m_{30} = 0$.

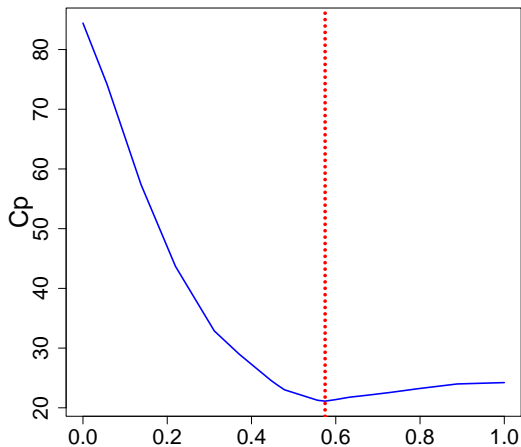
We choose λ by minimizing the estimated risk.

SpAM yields 6 nonzero functions. It correctly reports that $\hat{m}_{11} = \cdots = \hat{m}_{30} = 0$.

L_2 norms of fitted functions versus $1/\lambda$



Estimated Risk Versus λ



Summary for today

- Smoothing methods compute local averages, weighting points by a kernel
- The details of the kernel don't matter much
- Mercer kernels using penalization rather than smoothing
- Defining property: Matrix \mathbb{K} is always positive-definite (non-negative eigenvalues)
- Equivalent to a type of ridge regression in function space
- The curse of dimensionality limits use of both approaches to low dimensions
- A compromise between nonparametric and linear models is to use additive models