IEOR 265 – Lecture 9
Other Models

1 Single-Index Model

Recall the following single-index model

$$y_i = g(x_i'\beta) + \epsilon_i,$$

where $y_i \in \mathbb{R}$, $x_i, \beta \in \mathbb{R}^p$, $g(\cdot)$ is an unknown nonlinear function, and ϵ_i are noise. The data x_i are i.i.d., and the noise has conditionally zero mean $\mathbb{E}[\epsilon_i | x_i] = 0$. Such single-index models can be used for asset pricing, and here the $g(\cdot)$ can be thought of as an infinite-dimensional nuisance parameter.

The difficulty in this situation is that the function $g : \mathbb{R} \to \mathbb{R}$ is unknown, because otherwise we could simply use nonlinear least squares

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - g(x_i'\beta))^2$$

to provide an estimate of β. Because $g(\cdot)$ is not known, one possible approach is to begin with nonlinear least squares and replace $g(\cdot)$ with something that looks similar. If we define

$$\hat{\beta} = \arg\min_{\eta} \sum_{i=1}^{n} (y_i - \mathbb{E}[y_i | z = x_i'\eta])^2 = \arg\min_{\eta} \sum_{i=1}^{n} (y_i - \mathbb{E}[g(x_i'\beta) | z = x_i'\eta])^2$$

Now we do not know $\mathbb{E}[g(x_i'\beta) | z = x_i'\eta]$, but we could estimate this using the trimmed version of Nadaraya-Watson. (Strictly speaking, the standard estimator uses a leave-one-out version of the Nadaraya-Watson estimator in which the i-th data is excluded when estimating at i, but this change does not affect the convergence results.)

One reason this estimator is interesting is that the objective of the defining optimization involves an estimate itself, and it is an example of an adaptive optimization problem. This approach cannot always be used, but it turns out that the Nadaraya-Watson estimator is a good choice for doing so. This will be discussed later in the course.

2 Linear Support Vector Machine

Consider the following nonlinear model:

$$y_i = \text{sign}(x_i'\beta + \beta_0) + \epsilon_i,$$

where $y_i \in \{-1, 1\}$, $x_i, \beta \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$, and ϵ_i is noise. We can think of the y_i as labels of each x_i, and the ϵ_i noise represents (potential) mislabeling of each x_i. The intuitive picture is that $x_i'\beta + \beta_0 = 0$ defines a hyperplane that separates $\mathbb{R}^p$ into two parts, in which points on one side of the hyperplane are labeled 1 while points on the other side are -1. Also note that there is a normalization question because $x_i'(\gamma \cdot \beta) + \beta_0 = 0$ defines the same hyperplane for any $\gamma > 0$.

The key to identifying models is to observe that the boundaries of the half-spaces

$$\begin{aligned} x_i'\beta + \beta_0 &\leq -1 \\ x_i'\beta + \beta_0 &\geq 1 \end{aligned}$$

are parallel to the separating hyperplane $x_i'\beta + \beta_0 = 0$, and the distance between these two half-spaces is $2/\|\beta\|$. So by appropriately choosing β, we can make their boundaries arbitrarily close (or far) to $x_i'\beta + \beta_0 = 0$. Observe that for noiseless data, we would have

$$y_i(x_i'\beta + \beta_0) \geq 1.$$

So we could define our estimate by the coefficients that maximize the distance between the half-spaces (which is equivalent to minimizing $\|\hat{\beta}\|$ since the distance is $2/\|\hat{\beta}\|$). This would be given by

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \arg\min_{\beta_0, \beta} \ & \|\beta\|^2 \\ \text{s.t. } & y_i(x_i'\beta + \beta_0) \geq 1, \forall i. \end{aligned}$$

This formulation assumes there is no noise. But if there is noise, then we could modify the optimization to

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \arg\min_{\beta_0, \beta} \ & \|\beta\|^2 + \lambda \sum_{i=1}^{n} u_i \\ \text{s.t. } & y_i(x_i'\beta + \beta_0) \geq 1 - u_i, \forall i \\ & u_i \geq 0 \end{aligned}$$

The interpretation is that u_i captures errors in labels. If there are no errors, then $u_i = 0$, and if there is high error then u_i is high. The term $\lambda \sum_{i=1}^{n} u_i$ denotes that we would like to pick our parameters to tradeoff maximizing the distance between half-spaces with minimizing the amount of errors.

3 Summary

We have covered several regression techniques:

Model	Possible Techniques
Linear	Ordinary Least Squares (OLS)
Nonlinear with Known Form	Nonlinear Least Squares (NLS)
Nonlinear with Unknown Form	Nadaraya-Watson (NW) Local Linear Regression (LLR)
Partially Linear Model	Procedure with NW + OLS
Single-Index Model	Procedure with NW + OLS
Linear Classification	Linear Support Vector Machine (SVM)

We have also covered several types of abstract structure:

Structure	Possible Regularizations
Collinearity/Manifold Structure	L2 Regularization/Ridge Regression Exterior Derivative Estimator (EDE)
Sparsity	L1 Regularization/Lasso Regression
Group Sparsity	Sparse Group Lasso Regression

We can combine these regression technique and abstract structures for different scenarios. OLS, linear SVM, and LLR can have no special structure, collinearity/manifold structure, sparsity, group sparsity, or any combination of these. NW implicitly exploits collinearity/manifold structure, and L2 regularization instead acts as another form of trimming. NLS can have no special structure, sparsity, group sparsity, or any combination of these; however, L2 regularization can be used with NLS to provide shrinkage for the purposes of improving the bias-variance tradeoff.