IEOR 165 – Lecture Notes
Regularization

1 Maximum A Posteriori (MAP) Estimation

The MLE framework consisted of formulating an optimization problem in which the objective was the likelihood (as parametrized by the unknown model parameters) of the measured data, and the minimizer of the optimization problem gave our estimate. This is not the only framework for estimation, and it is interesting to consider alternative approaches.

An alternative approach uses a Bayesian framework. This approach begins similarly to MLE: Suppose we have a parameterized pdf $f(X_1, \ldots, X_n|\theta)$ and iid measurements X_i, for $i = 1, \ldots, n$, from this parameterized distribution. The new element of this framework is that we also assume that there is a probability distribution of the unknown parameters. So for instance, we have a pdf $g(\theta)$ for the unknown parameter. From a conceptual standpoint, the interpretation is we have some prior knowledge about what the possible values of the unknown parameters might be.

The pdf of the unknown parameter $g(\theta)$ is known as a prior distribution because it represents our knowledge about the model parameters without knowing any data. Now the idea is that because we have data X_i, we can use this to update our knowledge about the model parameters. In particular, we can define the posterior distribution

$$f(\theta|X_1, \ldots, X_n) = \frac{f(X_1, \ldots, X_n|\theta)g(\theta)}{\int_{u\in\Theta} f(X_1, \ldots, X_n|u)g(u)},$$

which describes our knowledge about the model parameters θ conditioned on the data we have observed. This equation is merely an application of Bayes's Theorem.

One idea to estimate the parameters is to compute the maximum *a posteriori* (MAP) estimate, which is defined by the maximizer to the following optimization problem

$$\max \big\{ f(\theta|X_1, \ldots, X_n) \;\big|\; \theta \in \Theta \big\},$$

and the interpretation is that our estimate is the value that maximizes the posterior likelihood (of the parameter given the data measured). One interesting observation that is useful in practice is to note that the denominator $\int_{u\in\Theta} f(X_1, \ldots, X_n|u)g(u)$ in the equation defining the posterior distribution is a constant. As a result, we can equivalently compute the MAP estimate by solving

$$\max \big\{ f(X_1, \ldots, X_n|\theta)g(\theta) \;\big|\; \theta \in \Theta \big\}.$$

This is a considerable simplification because computing the integral $\int_{u\in\Theta} f(X_1, \ldots, X_n|u)g(u)$ can be very hard in general. A second simplification is that because we have assumed the X_i are

iid, if $f(X_i|\theta)$ is the conditional pdf of X_i given θ then we have that

$$f(X_1, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta).$$

This means the MAP can be computed by solving

$$\max \big\{ \textstyle\prod_{i=1}^{n} f(X_i|\theta) \cdot g(\theta) \;\big|\; \theta \in \Theta \big\}.$$

A final simplifcation is that we can take the negative logarithm, just as we did for the MLE. This means the MAP can be computed by solving

$$\min \big\{ -\log g(\theta) - \textstyle\sum_{i=1}^{n} \log f(X_i|\theta) \;\big|\; \theta \in \Theta \big\}.$$

This can be a considerable simplification when computing derivatives or gradients.

1.1 Example: Least Squares with Gaussian Prior

Consider the linear model

$$Y_i = x_i^\mathsf{T} \beta + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are iid with *known* σ^2. In this model, we are assuming the x_i for $i = 1, \ldots, n$ are deterministic vectors (i.e., they are not random variables). Assume the prior distribution of the coefficients are iid Gaussians:

$$\beta_k \sim \mathcal{N}(0, 1/(2\lambda)).$$

Q: Write the MAP estimate as a quadratic program (QP).

A: We begin by noting the MAP estimate is given by

$$\hat{\beta} = \arg\max \exp\big(\textstyle\sum_{i=1}^{n} -(Y_i - x_i^\mathsf{T}\beta)^2/(2\sigma^2) \big) \prod_{k=1}^{p} \exp\big(-\lambda \beta_k^2 \big).$$

Using the same trick as in MLE, we instead minimize the negative logarithm:

$$\hat{\beta} = \arg\min \textstyle\sum_{i=1}^{n} (Y_i - x_i^\mathsf{T}\beta)^2/(2\sigma^2) + \sum_{k=1}^{p} \lambda \beta_k^2.$$

And this is the same as solving

$$\hat{\beta} = \arg\min \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \tilde{\lambda}\|\beta\|_2^2,$$

where $\tilde{\lambda} = 2\lambda\sigma^2$ and where matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ and vector $\mathbb{Y} \in \mathbb{R}^n$ are such that the i-th row of $\mathbb{X}$ is x_i^T and the i-th row of $\mathbb{Y}$ is Y_i. This is a QP because it can be rewritten as

$$\hat{\beta} = \arg\min \beta^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X} + \tilde{\lambda}\mathbb{I})\beta - 2\mathbb{Y}^\mathsf{T}\mathbb{X}\beta.$$

1.2 Example: Least Squares with Laplacian Prior

We begin with the same linear model

$$Y_i = x_i^\mathsf{T}\beta + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are iid. Now instead assume the prior distribution of the coefficients are iid Laplacians with zero mean and scale parameter $1/\lambda$. Recall that the pdf of the Laplacian distribution with these parameters is

$$f(u) = \frac{\lambda}{2}\exp\big(-\lambda|u|\big).$$

Q: Write the MAP estimate as a quadratic program (QP).

A: We begin by noting the MAP estimate is given by

$$\hat{\beta} = \arg\max \exp\big(\textstyle\sum_{i=1}^n -(Y_i - x_i^\mathsf{T}\beta)^2/(2\sigma^2)\big)\prod_{k=1}^p \exp\big(-\lambda|\beta_k|\big).$$

Using the same trick as in MLE, we instead minimize the negative logarithm:

$$\hat{\beta} = \arg\min \textstyle\sum_{i=1}^n (Y_i - x_i^\mathsf{T}\beta)^2/(2\sigma^2) + \sum_{k=1}^p \lambda|\beta_k|.$$

And this is the same as solving

$$\hat{\beta} = \arg\min \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \tilde{\lambda}\|\beta\|_1,$$

where $\tilde{\lambda} = 2\lambda\sigma^2$. This is a QP because it can be rewritten as

$$\begin{aligned}\hat{\beta} = \arg\min\ & \beta^\mathsf{T}\mathbb{X}^\mathsf{T}\mathbb{X}\beta - 2\mathbb{Y}^\mathsf{T}\mathbb{X}\beta + \tilde{\lambda}\cdot t \\ \text{s.t.}\ & -m_k \le \beta_k \le m_k, \qquad \forall k = 1,\dots,p \\ & \textstyle\sum_{k=1}^p m_k \le t.\end{aligned}$$

2 Regularization

Regularization is the process of purposely increasing the bias of an estimator so that the overall estimation error decreases. There are different types of regularization that are available to us depending on what underlying structure may be present in the system that we are trying to model. Incorporating appropriate regularization is arguably the most important step in modeling, and the key problem is to understand what kinds of additional structure can be imposed onto the model.

3 L2 Regularization

We have already seen the example of L2 regularization imposed on the OLS method:

$$\hat{\beta} = \arg\min \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

As discussed in the previous lecture, this approach can be interpreted as providing (proportional) statistical shrinkage. This shrinks the estimated coefficients of the linear model towards zero, in exchange for a reduction in variance of the estimated coefficients. As shown above, L2 regularization for the OLS model also has a Bayesian interpretation.

3.1 Example: Collinearity

Collinearity is an issue that frequently arises when trying to construct linear models from data. Suppose we would like to build a linear model that has an output of *diastolic blood pressure* and inputs of

- age
- salary
- cholesterol level
- weight
- height
- body fat percentage

Constructing this model is more difficult than it might initially seem. The reason is that the inputs are expected to display collinearity, meaning that the inputs themselves are correlated to each other. For instance, body fat percentage will be correlated to weight and height. Similarly, cholesterol level will be correlated to age and weight.

It turns out that this kind of structure causes the estimation error of OLS to increase significantly. The reasons are technical but can be interpreted either in the language of differential geometry or in the context of numerical conditioning. Given the challenges faced by such collinearity, it is interesting to consider how to improve estimation error of OLS. One approach that works well in practice is to incorporate L2 regularization into OLS.

4 L1 Regularization

Another type of regularization is known as L1 regularization, and it consists of solving the following optimization problem

$$\hat{\beta} = \arg\min \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1,$$

where λ is a tuning parameter. This approach is also called Lasso regression, and is a popular technique in machine learning. As shown above, this method has a Bayesian interpretation. It also has a "bias-variance tradeoff" interpretation.

4.1 Soft-Thresholding Interpretation

The Lasso regression estimate has an important interpretation in the bias-variance context. For simplicity, consider the one-dimensional case in which $\sum_{i=1}^n X_i^2 = 1$. In this case, the objective of the Lasso regression is

$$\begin{aligned}
\hat{\beta} &= \arg\min \sum_{i=1}^n (Y_i - X_i \cdot \beta)^2 + \lambda|\beta| \\
&= \arg\min \sum_{i=1}^n Y_i^2 - 2Y_i X_i \beta + X_i^2 \beta^2 + \lambda|\beta| \\
&= \arg\min(\textstyle\sum_{i=1}^n -2Y_i X_i)\beta + (\sum_{i=1}^n X_i^2)\beta^2 + \lambda|\beta| + \sum_{i=1}^n Y_i^2 \\
&= \arg\min(\textstyle\sum_{i=1}^n -2Y_i X_i)\beta + \beta^2 + \lambda|\beta| + \sum_{i=1}^n Y_i^2.
\end{aligned}$$

Note that even though the objective is not differentiable, we can break the problem into three cases. In the first case, $\beta > 0$ and so setting the derivative equal to zero gives

$$2\beta + (\textstyle\sum_{i=1}^n -2Y_i X_i) + \lambda = 0 \Rightarrow \hat{\beta} = \sum_{i=1}^n Y_i X_i - \lambda/2.$$

In the second case, $\beta < 0$ and so setting the derivative equal to zero gives

$$2\beta + (\textstyle\sum_{i=1}^n -2Y_i X_i) - \lambda = 0 \Rightarrow \hat{\beta} = \sum_{i=1}^n Y_i X_i + \lambda/2.$$

In the third case, $\beta = 0$.

For reference, we also compute the OLS solution in this special case. If we define $\hat{\beta}_{\mathsf{OLS}}$ to be the OLS solution, then a similar calculation to the one shown above gives that $\hat{\beta}_{\mathsf{OLS}} = \sum_{i=1}^n Y_i X_i$. And so comparing the OLS solution to the Lasso regression solution, we have that

$$\hat{\beta} = \begin{cases} \hat{\beta}_{\mathsf{OLS}} + \lambda/2, & \text{if } \hat{\beta} = \hat{\beta}_{\mathsf{OLS}} + \lambda/2 < 0 \\ \hat{\beta}_{\mathsf{OLS}} - \lambda/2, & \text{if } \hat{\beta} = \hat{\beta}_{\mathsf{OLS}} - \lambda/2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

This can be interpreted as a soft thresholding phenomenon, and it is another approach to balancing the bias-variance tradeoff.

4.2 Dual Formulation

Consider the following optimization problem

$$\hat{\beta} = \arg\min_{\beta} \{\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 : \phi(\beta) \leq t\},$$

where $\phi : \mathbb{R}^p \to \mathbb{R}$ is a penalty function with the properties that it is convex, continuous, $\phi(0) = 0$, and $\phi(u) > 0$ for $u \neq 0$. It turns out that there exists λ such that the minimizer to the above optimization is identical to the minimizer of the following optimization

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda\phi(\beta).$$

This relationship can be established using duality theory from optimization, but that material is covered in other courses.

However, this result is useful because it has a graphical interpretation that provides additional insight. Visualizing the constrained form of the estimator provides intuition into why L2 regularization does not lead to sparsity, whereas L1 regularization does. By sparsity, we mean that most of the estimated linear coefficients β will be equal to zero. This is important because zero coefficients indicate no relationship between the corresponding input variables and the output variable. Furthermore, in many applications we can have an extremely large number of inputs (several thousand inputs is not uncommon), and we may not know *a priori* which inputs are relevant. The Lasso regression method is one approach that helps choose relevant input variables, and is popular in machine learning for this reason.

5 Elastic Net

In some situations, we might have collinearity of the input variables *and* a large number of input variables. In these cases, we can combine L1 and L2 regularization and solve

$$\hat{\beta} = \arg\min \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 + \mu\|\beta\|_1.$$

This method is known as the *elastic net*.