Linear regression

Notation

We will use operator notation. Typically, P is used to represent expectation over distribution p:

$$Pf(X) = \int f(x)p(x)dx.$$

For example, if we use $\mathbb{P}_n$ to denote empirical density, then

$$\mathbb{P}_n f(X) = \int f(x)\frac{1}{N}\sum_i \delta(x - x_i)dx = \frac{1}{N}\sum_i f(x_i).$$

We will use $\mathbb{E}$ to represent average over all the randomness.

X is usually denoted as random variable (predictors) and $\mathbb{X}$ is denoted as design matrix.

Least square estimation

Mean square error and least square solution

The least square estimatiton is a decision criterion based on L2 risk (mean square error):

$$f^* = \arg\min_f P(Y-f(X))^2,$$

where X is predictor and Y is response. This criterion may stem from the MLE based on Gaussian conditional mean assumption, but anyway we will use this criterion without any assumption. By conditioning on X, we can find the best estimation that minimizes L2 risk is exactly conditional mean (see Appendix-1):

$$f^*(X) = P(Y|X).$$

In linear regression, we restrict to linear functions $f(x)=\beta^T x$, so we only search throguh linear function space:

$$\beta^* = \arg\min_\beta P(Y-\beta^T X)^2.$$

Assume PXX^T is positive definite (thus revertible), then the problem above is a convex optimization and we easily find β^* by letting its gradient be zero:

$$PXX^T\beta^* = P(XY) \Rightarrow \beta^* = (PXX^T)^{-1}PXY.$$

This is also called the \textbf{normal equation}. Essentially $\hat{Y} = (\beta^*)^T X$ is the \textbf{best linear estimator} we can find. It is shown that $(\beta^*)^T X$ is actually the projection of conditional mean $f^*(X) = P(Y|X)$ onto linear function space (see Appendix-2).

Here comes the problem. Since we do not know distribution P, so in fact we can only estimate β^*. The estimation $\hat{\beta}_n$ is defined by empirical risk minimizer (ERM):

$$\hat{\beta}_n = \arg\min_\beta \mathbb{P}_n(Y-\beta^T X)^2.$$

Similarly, we obtain the normal equation for empirical least square estimation:

$$\mathbb{P}_n XX^T\hat{\beta}_n = \mathbb{P}_n XY \Rightarrow \hat{\beta}_n = (\mathbb{P}_\mathrm{n} XX^T)^{-1}\mathbb{P}_n XY.$$

Best model as projection

There are some interesting facts we should know about the relations between regression function $P(Y|X)$ and best linear model $(\beta^*)^T X$.

```
\begin{remark}
The best linear model $(\beta^*)^TX$ is the **projection** of regression function $P(Y|X)$ onto linear function space. In ge
\end{remark}
```

```
\begin{proof}
Note that we can decompose L2 risk into
\begin{equation}
P(Y-f(X))^2=P(Y-P(Y|X)+P(Y|X)-f(X))^2=P\text{Var}(Y|X)+P(P(Y|X)-f(X))^2.
\end{equation}
The first term has nothing to do with hypothesis space $F$. So our best model in terms of L2 risk is
\begin{equation}
f^*=\arg\min_{f\in F}P(Y-f(X))^2=\arg\min_{f\in F}P(P(Y|X)-f(X))^2.
\end{equation}
It suggests we are actually projecting $P(Y|X)$ onto $F$ and get thet best model.
\end{proof}
```

Risk decomposition

The best model and $P(Y|X)$ are all independent of samples (data), and in practice the most ideal situation is that $\hat{\beta}_n \approx \beta^*$ (we cannot find $P(Y|X)$). To study the error of given $\hat{\beta}_n$, we can decompose its error into several parts:

$$\begin{aligned} E(Y - \hat{\beta}_n^T X)^2 &= P\mathrm{Var}(Y|X) + E(P(Y|X) - \hat{\beta}_n^T X)^2 = \mathrm{Noise} + E(P(Y|X) - (\beta^*)^T X + (\beta^*)^T \\ &= \mathrm{Noise} + P(P(Y|X) - (\beta^*)^T X)^2 + E((\beta^* - \hat{\beta}_n)^T X)^2 = \mathrm{Noise} + \mathrm{Appr.\ Err} + \end{aligned}$$

where $\mathrm{Noise} = P\mathrm{Var}(Y|X)$ refers to noise in data, $\mathrm{Appr.\ Err} = P(P(Y|X) - (\beta^*)^T X)^2$ is called **approximation error** that represents risk of our best model and $\mathrm{Est.\ Err} = P((\beta^* - \hat{\beta}_n)^T X)^2$ is called **estimation error** that reflects the difference between our estimation and the best model. Only estimation error depends on our samples. The cross term is zero due to normal equation.

A typical trade-off here is between approximation error and estimation error. It depends on the size (complexity) of our hypothesis space. Consider two extreme cases: if we allow all kinds of possible functions, then definitely the approximation error would be zero, while the estimation error would be huge; on the contrary, if we only allow the simplest function (say, pre-determined constant prediction), then there would not be any estimation error, but the approximation error is large. In principle, estimation error would shrink as sample size grows, so a larger sample size allows a large hypothesis space.

Trade-off between variance and bias

Notice that estimation error can be further decomposed into the following:

$$E(X^T(\beta^* - \hat{\beta}_n))^2 = P(X^T E(\beta^* - \hat{\beta}_n)(\beta^* - \hat{\beta}_n)^T X).$$

Here $E(\beta^* - \hat{\beta}_n)(\beta^* - \hat{\beta}_n)^T$ is called **mean square error** of parameter estimation. It can be decomposed into bias and variance:

$$E(\beta^* - \hat{\beta}_n)(\beta^* - \hat{\beta}_n)^T = Var(\hat{\beta}_n) + (E\hat{\beta}_n - \beta^*)(E\hat{\beta}_n - \beta^*)^T.$$

As we can see latter, the idea of shrinkage method (regularization) is to increase bias a little but reduce variance a lot, which leads to the overall shrink of estimation error.

Asymptotic properties of ordinary least square

In this section, we are going to show that least square solution $\hat{\beta}_n$will converge to best linear estimator β^*. Based on its asympotical behavior, we can construct confidence interval for β^*.

To show this, consider the difference between them:

$$\begin{aligned}\hat{\beta}_n - \beta^* &= (\mathbb{P}_\mathrm{n} XX^T)^{-1}\mathbb{P}_\mathrm{n} XY - \beta^* = (\mathbb{P}_\mathrm{n} XX^T)^{-1}\mathbb{P}_\mathrm{n} X(Y - X^T\beta^*) \\ &= (\mathbb{P}_\mathrm{n} XX^T)^{-1}(\mathbb{P}_\mathrm{n} - P)X(Y - X^T\beta^*).\end{aligned}$$

The last line is due to normal equation $PX(Y - X^T\beta^*) = 0$. Now according to law of large number and central limit theorem,

$$\begin{gathered}\mathbb{P}_\mathrm{n} XX^T \to PXX^T \\ \sqrt{n}(\mathbb{P}_\mathrm{n} - P)X(Y - X^T\beta^*) \rightsquigarrow N(0, PXX^T(Y - X^T\beta^*)^2).\end{gathered}$$

So $\hat{\beta}_n \to \beta^*$ with

$$\sqrt{n}(\hat{\beta} - \beta^*) \rightsquigarrow N(0, \Sigma),$$

where $\Sigma = (PXX^T)^{-1}PXX^T(Y - X^T\beta^*)^2(PXX^T)^{-1}$. In practice, we can estimate Σ using $\hat{\Sigma} = (\mathbb{P}_\mathrm{n} XX^T)^{-1}\mathbb{P}_\mathrm{n} XX^T(Y - X^T\hat{\beta}_n)^2(\mathbb{P}_\mathrm{n} XX^T)^{-1}$

Ridge regression

Notice that if linear model is correct, namely $P(Y|X) = X^T\beta^*$, then least square estimator $\hat{\beta}_n$ is unbiased because:

$$E\hat{\beta}_n = E(P_nXX)^{-1}P_nXY = E_X(P_nXX^T)^{-1}P_nXE_{Y|X}Y = E_X(P_nXX^T)^{-1}P_nXX^T\beta^* = \beta^*.$$

But as we can see in trade-off between variance and bias, it is not always a good idea to make a unbiased estimation if we want MSE to be lower. The idea of regularization is to significantly reduce estimation error through variance/bias trade-off, though approximation error may increase a little.

This motivates ridge regression. Ridge regression is defined as the following modified optimization problem:

$$\hat{\beta}_n = \arg\min_\beta \mathbb{P}_\mathrm{n}(Y - X^T\beta)^2 + \lambda\|\beta\|_2^2.$$

It can be shown that this is definitely a convex problem given a positive λ. So the closed form solution for ridge regression is obtained by setting the gradient to zero:

$$\hat{\beta}_n = (\mathbb{P}_\mathrm{n} XX^T + \lambda I)^{-1}\mathbb{P}_\mathrm{n} XY.$$

There are plenty of ways to explain ridge regression.

Ridge regression as constrained optimization

We can prove that ridge regression (???) is equivalent to the following constrained optimization with certain $\lambda = \lambda(C)$:

$$\hat{\beta}_n = \arg\min_{\beta} \mathbb{P}_\mathrm{n}(Y - X^T\beta)^2,$$
$$\text{s.t. } \|\beta\|_2^2 \leq C.$$

This can be proved by using KKT conditions (see Appendix-3). It basically says that when doing ridge regression, we are actually doing least square but in a constrained hypothesis set. So intuitively the approximation error would be larger than OLS, since we have a smaller hypothesis space, but the estimation error would be (maybe) reduced. It had be shown that there exists some λ such that the mean square error of ridge is stricly less than OLS:

$$E\left\|\hat{\beta}_n^{\text{ridge}} - \beta^*\right\|_2^2 \leq E\left\|\hat{\beta}_n^{\text{OLS}} - \beta^*\right\|_2^2$$

Ridge regression as MAP

Another way to explain ridge is through Bayesian estimation. Suppose linear model is correct, that is $Y \sim N(X^T\beta, \sigma^2)$ and $\beta \sim N(0, \tau^2)$. To make a point estimation of β, we decide to maximize a posterior probability (MAP), that is

$$\hat{\beta}_n = \arg\min -\log p(\beta|\mathbb{X}, \mathbb{Y}) = \arg\min -\log p(\mathbb{Y}|\beta, \mathbb{X}) - \log p(\beta) = \arg\min_{\beta} \mathbb{P}_\mathrm{n}(Y - X^T\beta)^2 +$$

So we can define $\sigma^2/n\tau^2 \triangleq \lambda$ as the regularization paramter. In practice, it is useful to first estimate $\hat{\sigma}^2/n\hat{\tau}^2$ as a initial value for λ. So in this point of view, ridge regression is just OLS plus the prior knowledge that each component of paramters β^* should be small and close to zero. This is why as a result we would shrink $\hat{\beta}_n$.

Ridge regression as weighted projection

Recall that least square prediction $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}_n^{OLS}$, in terms of design matrix $\mathbb{X}$ and response $\mathbb{Y}$, can be viewed as orthogonal projection of $\mathbb{Y}$ onto range of $\mathbb{X}$:

$$\hat{\mathbb{Y}}^{OLS} = \mathbb{X}(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}^T\mathbb{Y} = UU^T\mathbb{Y} = \sum_j u_j u_j^T \mathbb{Y},$$

where $\mathbb{X} = UDV^T$ is the SVD of design matrix. If we check the closed form solution of ridge, we find:

$$\hat{\mathbb{Y}}^{ridge} = \mathbb{X}(\mathbb{X}\mathbb{X}^T + \lambda I)^{-1}\mathbb{X}^T\mathbb{Y} = \sum_j u_j \mathrm{diag}\{\frac{d_i^2}{d_i^2 + \lambda}\} u_j^T \mathbb{Y}.$$

So now we can clearly see how ridge regression shrinks the coefficients. Roughly speaking, the degree of shrinkage depends on the sample variance of features (or say principal components): if variance of certain is small, then we think it is unimportant and shrink it a lot.

Ridge regression as dropout

Dropout, that is to randomly zero out some components of predictors, is a common trick for regularization. We will show that dropout is equivalent to ridge regression. Consider the following formal definition of least square estimation with dropout:

$$\hat{\beta}_n = \arg\min_{\beta} \sum_k \sum_n (Y_n - (X_n \cdot Z_k)^T \beta/(1-\phi))^2,$$

here Z_k are i.i.d. Bernulli vectors with $P(Z_{k,i} = 1) = 1 - \phi$. $X_n \cdot Z_k \triangleq (X_{n,1}Z_{k,1}, \ldots, X_{n,p}Z_{k,p})$. Basically we create much more data by dropout. As we can create as much as dropout data we want, so

$$\hat{\beta}_n \approx \arg\min_{\beta} E_Z \mathbb{P}_\mathrm{n} (Y - (X \cdot Z)^T \beta/(1-\phi))^2,$$

Solving this optimization problem gives exactly the ridge regression

$$\hat{\beta}_n = (\mathbb{P}_\mathrm{n} XX^T + \frac{\phi}{1-\phi}\mathrm{diag}\{\mathbb{P}_\mathrm{n} X_i^2\})^{-1} \mathbb{P}_\mathrm{n} XY.$$

The proof is shown in Appendix-4. Suppose each feature X_i is standardized, then dropout is the same as ridge regression in the sense that $\lambda = \phi/(1-\phi)$.

Ridge regression as noises addition

Just as we can create data in case of dropout, we can also create data by adding noisy predictors. Let $W \sim N(0, \tau^2 I)$, then the noisy optimization problem becomes

$$\hat{\beta}_n = \arg\min_{\beta} E_W \mathbb{P}_\mathrm{n} \big(Y - (X+W)^T \beta\big)^2.$$

By settin the gradient to zero, we find

$$\hat{\beta}_n = \big(\mathbb{P}_\mathrm{n} XX^T + \tau^2 I\big)^{-1} \mathbb{P}_\mathrm{n} XY.$$

So our regularization parameter λ can be seen as the magnitude of noise τ^2.

Lasso

Another shrinkage method is to apply L1 regularization:

$$\hat{\beta}_n = \arg\min_{\beta} \mathbb{P}_\mathrm{n} (Y - X^T \beta)^2 + \lambda \|\beta\|_1.$$

Sparse parttern of lasso solution

As we all knew, lasso leads to sparsity. To see this, let us consider $\mathbb{P}_\mathrm{n} XX^T = I$, then the lasso gives (see proof in Appendix-5)

$$\hat{\beta}_n^{\mathrm{lasso}} = max(0, \hat{\beta}_n - \lambda/2),$$

where $\hat{\beta}_n$ is the OLS solution. We can see that lasso solution is sparse in the sense that its j^{th} component is zero if $(\hat{\beta}_n)_j < \lambda/2$.

Lasso and robustness

Another way to explain lasso is through robust regression, which is defined by the following:

$$\hat{\beta}_n = \arg\min_{\beta} \max_{\Delta \in M} \|\mathbb{Y} - (\mathbb{X} + \Delta)\beta\|_2,$$

where M is a set of matrix. To obtain lasso, let us consider a specific M, that is

$$M = \left\{ \Delta \middle| \sqrt{\sum_i \Delta_{i,j}^2} \leq c_i \right\},$$

where $\{c_i\}$ is some constants. It can be shown that (see proof in Appendix-6)

$$\max_{\Delta \in M} \|\mathbb{Y} - (\mathbb{X} + \Delta)\beta\|_2 = \|\mathbb{Y} - \mathbb{X}\beta\|_2 + \sum_i c_i |\beta_i|.$$

So the robust regression becomes exactly lasso (actually, square root of lasso since) if we let $c_i = \lambda$. The hyperparameter λ thus be can thought of the perturbation of design matrix.

Appendix

Appendix-1

To see the conditional mean is the best estimator that minimizes L2 risk, we first condition on X and get

$$P(Y - f(X))^2 = P\left(P(Y - P(Y|X) + P(Y|X) - f(X))^2 |X\right) = P(Y - P(Y|X))^2 + P(P(Y|X)$$

then we can see that for each fixed X we should mimimize

$$f^*(X) = \arg\min_a (P(Y|X) - f(X))^2 \Rightarrow f^*(X) = P(Y|X).$$

Appendix-2

Recall the definition of $\beta^* = \arg\min_\beta P(Y - \beta^T X)$. It turns out

$$P(Y - \beta^T X)^2 = P(Y - f^*(X) + f^*(X) - \beta^T X)^2 = P(Y - f^*(X))^2 + P(f^*(X) - \beta^T X)^2.$$

Thus we have

$$\beta^* = arg \min_\beta P(f^*(X) - \beta^T X)^2,$$

which says $(\beta^*)^T X$ is the projection of conditional mean $f^*(X) = P(Y|X)$.

Appendix-3

The lagrangian of the constrained optimization (???) is

$$L(\beta, \lambda) = \mathbb{P}_{\mathrm{n}}(Y - X^T\beta)^2 + \lambda(\|\beta\|_2^2 - C),$$

where λ is the lagrangian multiplier. Suppose λ^* is optimal solution of dual problem, then from KKT conditions we know that optimal solution $\hat{\beta}_n$ should be the minimizer of $L(\beta, \lambda^*)$, which is

$$\hat{\beta}_n = \arg\min_{\beta} \mathbb{P}_{\mathrm{n}}(Y - X^T\beta)^2 + \lambda^*\|\beta\|_2^2.$$

For each C, we also solve a λ^*.

Appendix-4

Consider the gradient of dropout optimization:

$$\frac{\partial}{\partial\beta} E_z \mathbb{P}_{\mathrm{n}}(Y - (X \cdot Z)^T\beta/(1-\phi))^2 = \frac{1}{1-\phi} E_Z \mathbb{P}_{\mathrm{n}}(Y - (X \cdot Z)^T\beta/(1-\phi))X \cdot Z = 0.$$

Notice that $E_Z X \cdot Z = X(1-\phi)$ and $\mathbb{P}_{\mathrm{n}}(X \cdot Z)(X \cdot Z)^T = XX^T + \frac{\phi}{1-\phi}\mathrm{diag}\{X_i^2\}$, getting

$$\mathbb{P}_{\mathrm{n}} XY = \mathbb{P}_{\mathrm{n}} XX^T + \frac{\phi}{1-\phi}\mathrm{diag}\{\mathbb{P}_{\mathrm{n}} X_i^2\}\beta,$$

so

$$\hat{\beta}_n^{\mathrm{dropout}} = \left(\mathbb{P}_{\mathrm{n}} XX^T + \frac{\phi}{1-\phi}\mathrm{diag}\{\mathbb{P}_{\mathrm{n}} X_i^2\}\right)^{-1} \mathbb{P}_{\mathrm{n}} XY.$$

Appendix-5

Consider $\mathbb{P}_{\mathrm{n}} XX^T = I$. Then since

$$\mathbb{P}_{\mathrm{n}}(Y - X^T\beta)^2 = \mathbb{P}_{\mathrm{n}}(Y - X^T\hat{\beta}_n + X^T\hat{\beta}_n - X^T\beta)^2 = \mathbb{P}_{\mathrm{n}} Y^2 + \hat{\beta}_n^T\hat{\beta}_n + \left\|\hat{\beta}_n - \beta\right\|_2^2,$$

so the lasso becomes

$$\hat{\beta}_n^{\mathrm{lasso}} = \arg\min_{\beta} \left\|\hat{\beta}_n - \beta\right\|_2^2 + \lambda\|\beta\|_1.$$

This optimization problem can be solved by each component:

$$(\hat{\beta}_n^{\mathrm{lasso}})_j = \arg\min_{\beta} \left((\hat{\beta}_n)_j - \beta\right)^2 + \lambda\beta.$$

The gradient of objective function is

$$\frac{1}{2}\frac{\partial}{\partial\beta} = \beta - (\hat{\beta}_n)_j + \frac{\lambda}{2}\mathrm{sgn}(\beta).$$

The solution is simply

$$(\hat{\beta}_n^{\mathrm{lasso}})_j = \max(0, (\hat{\beta}_n)_j - \lambda/2).$$

Appendix-6

Our goal is to prove

$$\max_{\Delta \in M} \|\mathbb{Y} - (\mathbb{X} + \Delta)\beta\|_2 = \|\mathbb{Y} - \mathbb{X}\beta\|_2 + \sum_i c_i |\beta_i|$$

with $M = \left\{\Delta | \sqrt{\sum_i \Delta_{i,j}^2} \le c_j\right\}$. First note that from triangle inequality we have

$$\|\mathbb{Y} - (\mathbb{X} + \Delta)\beta\|_2 = \left\|\mathbb{Y} - \mathbb{X}\beta - \sum_i \Delta_i \beta_i\right\|_2 \le \|\mathbb{Y} - \mathbb{X}\beta\|_2 + \sum_i c_i |\beta_i|.$$

To show there exists a Δ such that triangle equality holds, we simply choose a Δ such that $\Delta\beta$ is in the same direction of $\mathbb{Y} - \mathbb{X}\beta$. Therefore we finish the proof.