18.650: Fundamentals of Statistics November 20, 2023

Lecture 26 — Model selection

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Motivation: avoiding overfitting

Consider the standard linear regression model $Y_i = X_i^T\beta + \epsilon_i$, $i = 1, \dots, n$, where (X_i, Y_i), $i = 1, \dots, n$ are the observed data, with $X_i \in \mathbb{R}^k$ and $Y_i \in \mathbb{R}$. Recall that $\hat{\beta}$ (aka $\hat{\beta}^{\mathrm{LS}}$) minimizes the sum of squared residuals:

$$\hat{\beta} = \mathrm{argmin}_{\beta\in\mathbb{R}^k} \sum_{i=1}^n (Y_i - X_i^T\beta)^2 = \mathrm{argmin}_{\beta\in\mathbb{R}^k} \|\vec{Y} - \mathbb{X}\beta\|^2.$$

Let us write out X_i in terms of its k coordinates: $X_i = (X_i^{(1)}, \dots, X_i^{(k)})$. The equation $Y_i = X_i^T\beta + \epsilon_i$ then takes the form

$$Y_i = \beta^{(1)}X_i^{(1)} + \dots + \beta^{(k)}X_i^{(k)} + \epsilon_i. \tag{1}$$

For example, Y_i could be patient i's blood pressure (bp), X_i could be
$X_i = (\mathrm{height}_i, \mathrm{weight}_i, \text{heart rate}_i, \mathrm{age}_i)$, and (1) would look like

$$\mathrm{bp}_i = \beta^{(1)} \cdot \mathrm{height}_i + \beta^{(2)} \cdot \mathrm{weight}_i + \beta^{(3)} \cdot \mathrm{heartrate}_i + \beta^{(4)} \cdot \mathrm{age}_i + \epsilon_i$$

1.1 Thought experiment: adding junk variables

Suppose we create s spurious variables which are totally irrelevant measurements of each patient, e.g. birthday, favorite color, house number. Let $X_i^{(k+1)}, \dots, X_i^{(k+s)}$ denote these measurements, which are essentially just noise. Call the new vector of $k + s$ measurements $X_i[k + s]$:

$$\begin{aligned} X_i &= (X_i^{(1)}, \dots, X_i^{(k)}) \\ &\to X_i[k+s] = (X_i^{(1)}, \dots, X_i^{(k)}, \underbrace{X_i^{(k+1)}, \dots, X_i^{(k+s)}}_{\text{noise}}) \end{aligned}$$

Now, find the least squares estimator for the new dataset with the extra measurements:

$$\hat{\beta}[k+s] = \mathrm{argmin}_{\beta\in\mathbb{R}^{k+s}} \sum_{i=1}^n (Y_i - X_i[k+s]^T\beta)^2 = \mathrm{argmin}_{\beta\in\mathbb{R}^{k+s}} \|\vec{Y} - \mathbb{X}[k+s]\beta\|^2.$$

As you might guess, $\hat{\beta}[k+s]$ leads to a better fit of the data, in the sense that the sum of squared residuals will decrease:

$$\left\|\vec{Y} - \mathbb{X}[k+s]\hat{\beta}[k+s]\right\|^2 \leq \left\|\vec{Y} - \mathbb{X}\hat{\beta}\right\|^2$$

See Figure 1 for visualization.

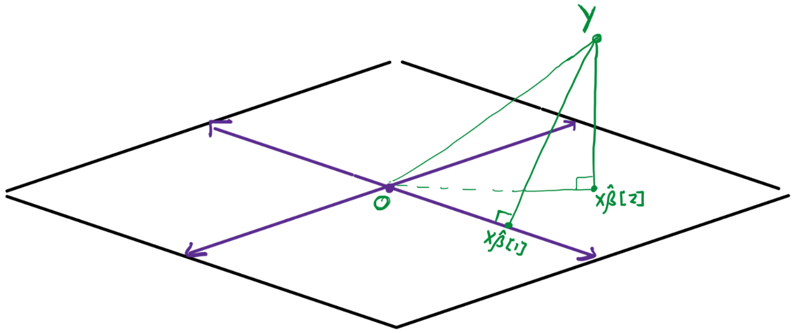

Figure 1: Increasing the number of features amounts to projecting onto a higher dimensional subspace. As you increase the dimension of the subspace, you decrease the codimension, and so the norm of the residual $\|Y - \mathbb{X}\hat{\beta}\|$ decreases.

And if you take s to be $n-k$ (so that the total number of variables is $k+s=n$), the sum of squared residuals will become exactly zero, leading to a perfect fit:

$$Y_i = X_i[n]\hat{\beta}[n], \quad \text{for all } i = 1, \ldots, n$$

exactly. This is because, as you increase the number of variables in the predictor

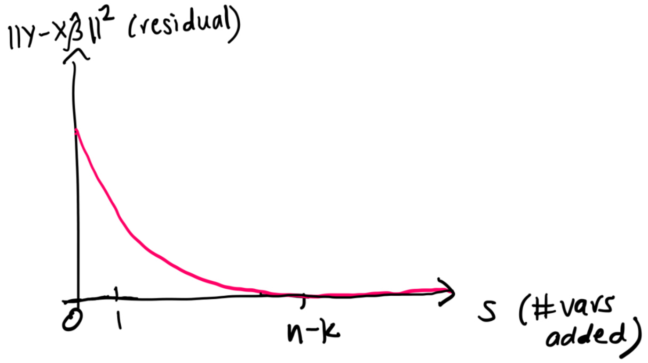

Figure 2: Norm of residual as a function of number of junk variables added.

X, you have more degrees of freedom to fit the response Y. Remember that $\mathbb{X}\hat{\beta}$ is the projection of $\vec{Y}$ onto a k dimensional subspace, and $\hat{\epsilon} = \vec{Y} - \mathbb{X}\hat{\beta}$ is the residual of dimension $n-k$. As you increase k, you can "explain" more of $\vec{Y}$ using $\mathbb{X}\hat{\beta}$, and the residual $\hat{\epsilon}$ will become smaller.

We see that

you can explain any Y perfectly with enough junk.

This is *not* a good thing because it leads to *overfitting*, which is when a model performs well on training data, but very poorly on unseen data.

2 Model selection

In our thought experiment, we started with a few variables k and saw what happened when you add s more junk variables.

Remark.

> In this lecture, the "variables" refer to the entries of the vector X.

But in real datasets, we typically have the opposite scenario. We start with too many variables — we don't know which ones might be relevant, so we include them all. For example, if we want to predict blood pressure from the genetic sequence, we might have $n = 1000$ patients and $k = 20,000$ genes for each patient. We then face the challenge of paring down the model. This is known as *model selection.*

The goal of model selection is to choose a subset $S \subset \{1, 2, \dots, k\}$ of the variables which matter most to explain Y. The model with variables S then takes the form

$$Y = \sum_{i \in S} \beta^{(i)} X^{(i)} + \epsilon.$$

2.1 Hypothesis testing for model selection?

We have already seen one way to do model selection: Wald's test of whether $\beta^{(j)}$ is nonzero, which corresponds to the variable $X^{(j)}$ being relevant to predict Y:

$$H_{0j} : \beta^{(j)} = 0 \quad \text{vs} \quad H_{1j} : \beta^{(j)} \neq 0.$$

If we run this test for each $j = 1, \dots, k$, then we are in the *multiple hypothesis testing* framework. This requires either

- Bonferroni correction — too conservative!

- BH — requires test statistics to be independent. Not true for linear regression!

Moreover:

- Both tests require asymptotics, $n \to \infty$. In fact, it's not enough for n to be large. Rather, the asymptotics kick in only when $n - k$ is large, because

  the effective sample size is $n - k$.

  If $n < k$, as in the genetics example, then our effective sample size is zero.

- If we start with a big model, what we really want to do is to test

  $$H_{0j} : \beta^{(j)} \neq 0 \quad \text{vs} \quad H_{1j} : \beta^{(j)} = 0.$$

  But the hypothesis testing framework does not allow us to do this.

2.2 A better framework

For each $S \subset \{1, \ldots, k\}$, we will define a "score" for each subset $S \subset \{1, 2, \ldots, k\}$, and then choose the S with the best score.

A bad choice would be to say the score is high if $\|\vec{Y} - \mathbb{X}\hat{\beta}\|^2$ is small. Indeed, we just saw that this leads to overfitting. Moreover, the quantity $\|\vec{Y} - \mathbb{X}\hat{\beta}\|^2$ depends on units — so we cannot really compare this score across different situations. To address this second issue, we want a unitless/dimensionless score between 0 and 1, like the p-value. This property is satisfied by R^2 defined below.

Definition 2.1: R-squared

The R^2 or coefficient of determination is a score between 0 and 1 that measures the fit of the model to the data.

$$R^2(S) = 1 - \frac{\|\vec{Y} - \mathbb{X}\hat{\beta}(S)\|^2}{\|\vec{Y} - \bar{Y}_n \mathbb{1}\|^2},$$

where $\mathbb{1}$ is the all ones vector, and $\hat{\beta}(S)$ is the least squares coefficient vector for the model using only the variables S.

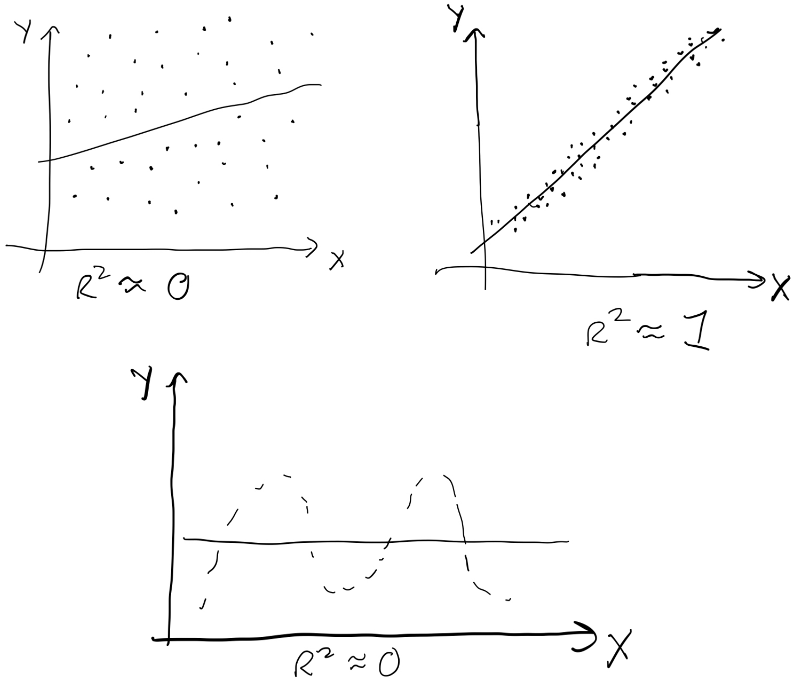

Figure 3: The R^2 measures the accuracy of the linear fit. When there is no relationship between X and Y or when the relationship is nonlinear, then R^2 is close to zero.

Remark.

> - Note that $R^2(S)$ is clearly less than 1.
> - The denominator is the sum of squared residuals for the best fit to the Y_i by a constant function. This is essentially using $f(x) = \mathbb{E}[Y \mid X] \approx \mathbb{E}[Y]$.
> - If $X_i^{(1)} = 1$, i.e. if the first variable in the X_i is the constant 1, then the least squares error $\|\vec{Y} - \mathbb{X}\hat{\beta}(S)\|^2$ is no larger than $\|\vec{Y} - \bar{Y}_n \mathbb{1}\|^2$, provided we include $1 \in S$. In this case, we can guarantee that $R^2(S) \geq 0$.
> - The larger the set S, the larger $R^2(S)$ will be.

The last point in the above remark has to do with the overfitting problem: we don't want to take $R^2(S)$ to be *too* large, because this is probably a sign we have entered overfitting territory. The next algorithm gives one way to choose the model based on $R^2(S)$.

Definition 2.2: Forward or greedy model selection

Initialize $S = \emptyset$.

Select $j_{\max}$ such that $R^2(S \cup \{j_{\max}\}) \geq R^2(S \cup \{j\})$ for all $j \in \{1, \ldots, k\} \setminus S$.

Set S to be $S = S \cup \{j_{\max}\}$.

Repeat until $R^2(S)$ plateaus, or once you first exceed some preset R^2 value, e.g. $R^2 = 0.8$.

In other words, this algorithm adds the variable that looks most promising at every step; see Figure 4.

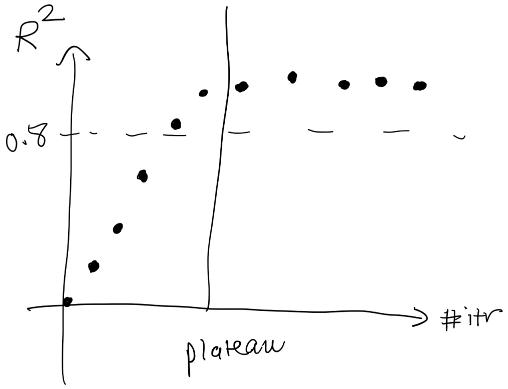

Figure 4: The greedy model selection algorithm stops when R^2 reaches some preset value or once $R2^(S)$ plateaus.

2.3 Information criteria: another common choice of score

The following are three common choices of score:

AIC: Akaike information criterion

$$\text{AIC} = 2|S| - 2\ell_n(\hat{\beta}(S))$$

BIC: Bayesian information criterion

$$\text{BIC} = (\log n)|S| - 2\ell_n(\hat{\beta}(S))$$

Mellow's C_p: equivalent to AIC in linear regression.

Here, $\hat{\beta}(S)$ is the MLE/least-squares estimator with model S, and ℓ_n is the log likelihood. The AIC and BIC trade off the size of the log likelihood with $|S|$.

We want the AIC/BIC to be *small*, by choosing $|S|$ as small as possible while preserving as large a value of $\ell_n(\hat{\beta}(S))$ as possible. The AIC/BIC curve as a function of $|S|$ will typically look like the one in Figure 5.

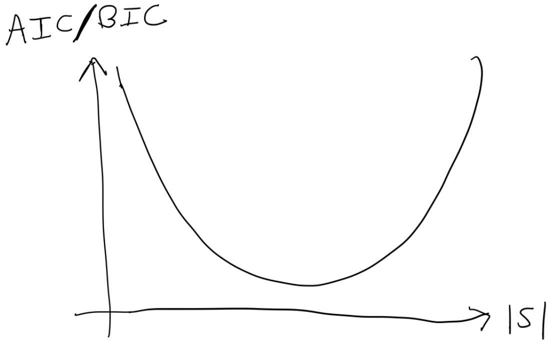

Figure 5: AIC and BIC as a function of $|S|$ (they are not literally identical to each other, but both AIC and BIC have roughly the same shape.

2.4 Computational issues

- For each value of $|S|$ there are $\binom{k}{|S|}$ models of size $|S|$. In total, there are 2^k possible models.

- Exhaustive search is only possible for very small k! Instead we resort to heuristics, such as the greedy forward algorithm described above. Note that this algorithm can be used with AIC/BIC in place of R^2.

- Another option is backward model selection: start with the full model, and drop the variable which leads to the smallest decrease of an IC.

- Yet another option is stepwise model selection: add a few variables at a time, then go back and see if removing some of them increases IC.

- An important remark: in contrast to using R^2 as the score, if we use the AIC/BIC we can just maximize the score without worrying about overfitting, since the ICs naturally penalize large models $|S|$.