6.S951 Modern Mathematical Statistics Fall 2024

Lecture 14 — November 5, 2024

Prof. Stephen Bates

Scribe: Sasha Voitovych, Anders Hoel

1 Outline

Agenda:

1. Predictive inference overview
   Key challenge: training residuals too small
2. Linear model calculation
3. Start conformal prediction

Recap:

1. CIs from CLT, Hoeffding
2. empirical CDFs, bands for DKW
3. bootstrap CIs
4. testing $\leftrightarrow$ CI Duality
5. Permutation tests

2 Prediction Inference

(Can think of it as confidence intervals in a supervised learning setting)

The setting. Let $\mathcal{X}$ be a space of covariates, and Y be the space of labels. For this lecture, we simply assume $\mathcal{X} = \mathbb{R}^d$ and $Y = \mathbb{R}$. We assume pairs (X_i, Y_i) are generated in i.i.d. manner from some distribution P, i.e.,

$$(X_i, Y_i) \overset{\text{iid}}{\sim} P, \text{ for } i = 1, \ldots, n.$$

Let $\mathcal{D}$ denote the training data, i.e.,

$$\mathcal{D} = \{(X_i, Y_i),\ i = 1, \ldots, n\}.$$

Our high level goal is to be able to predict Y's from X's. Specifically, we will assume there is a test point $X_{\text{test}}, Y_{\text{test}} \sim P$ generated from the same distribution independently of $\mathcal{D}$; we observe X_{test} and want to predict Y_{test}. More formally, we have the following structure.

Structure:

(i) Observe $\mathcal{D}$

(ii) Fit some model $\hat{f} : \mathcal{X} \to \mathcal{Y}$

(iii) Observe X_{test} and predict Y_{test}. Also, along with a prediction, we often aim to output a confidence interval for Y_{test}.

By producing a confidence interval (CI) in this setting we formally mean the following

$$C : \mathbb{R}^d \times \mathbb{R}^{n\times(d+1)} \to 2^{\mathbb{R}}.$$

In the above, $2^{\mathbb{R}}$ is a collection of all subsets of $\mathbb{R}$. I.e., C takes$X_{\text{test}} \in \mathbb{R}^d$ and $\mathcal{D} \in \mathbb{R}^{n\times(d+1)}$ as inputs, and produces a subset of $\mathbb{R}$. The coverage guarantee we are aiming for can be written as follows

$$P\left(Y_{\text{test}} \in C(X_{\text{test}}; \mathcal{D})\right) \geq 1 - \alpha, \quad \text{for some } \alpha \in (0, 1) \text{ fixed.} \tag{1}$$

We want this to hold as generally as possible i.e. for many distributions and many algorithms that produce $\hat{f}$. First, we consider a naive approach of achieving the above.

2.1 First attempt (wrong)

We can try to use residuals, defined as

$$r_i = Y_i - \hat{f}(X_i), \text{ for } i = 1, \ldots, n$$

to produce a confidence interval satisfying (1). In particular, let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ denote the $(\alpha/2)^{\text{th}}$ and $(1-\alpha/2)^{\text{th}}$ quantiles respectively of the collection $\{r_i\}_{i=1}^n$. Then, we can define a "naive" CI as

$$C^{\text{naive}}(X_{\text{test}}; \mathcal{D}) = (\hat{f}(X_{\text{test}}) + q_{\frac{\alpha}{2}}, \hat{f}(X_{\text{test}}) + q_{1-\frac{\alpha}{2}}))$$

The problem with the above is that r_i will often be too small, due to the phenomenon of *overfitting*. The model $\hat{f}$ will often "fit" the training dataset $\mathcal{D}$ too well, and, as a result, the residuals on $\mathcal{D}$ will be too "optimistic." For example, some modern machine learning algorithms produce overparameterized models for which $r_i = 0$ for all i, but we cannot expect such a model to fit the test set perfectly.

We will fix this problem using *data splitting*.

2.2 Second attempt (works)

We will split the dataset $\mathcal{D}$ in two parts and use first half to train $\hat{f}$ and the second to construct a CI. More, formally, we use (X_i, Y_i) $i = 1, \ldots, n/2$ to produce $\hat{f}$, and calculate the quantiles of the residuals of $\hat{f}$ only on the remaining points (X_i, Y_i) $i = n/2 + 1, \ldots, n$. This leads to a confidence interval of the form

$$C^{\text{split}}(X_{\text{test}}; \mathcal{D}) = (\hat{f}(X_{\text{test}}) + q_{\frac{\alpha}{2}}, \hat{f}(X_{\text{test}}) + q_{1-\frac{\alpha}{2}})),$$

where $q_{\alpha/2}, q_{1-\alpha/2}$ are quantiles of the collection $\{Y_i - \hat{f}(\hat{X}_i)\}$ for $i = n/2, \ldots, n$. Observe that $\hat{f}$ has not "seen" these datapoints. The following proposition will be proven in subsequent lectures as a corollary of a more general statement.

Proposition 1. *For the C^{split} defined above, we have*

$$P(Y_{test} \in C^{split}) \geq 1 - \alpha - \frac{1}{\frac{n}{2} + 1}$$

Note that the above works for any distribution and any model $\hat{f}$. Moreover, one can show an upper bound on the coverage as well, i.e., the coverage is in fact close to $1 - \alpha$.

3 Linear Model Calculation

To formally argue that the approach in Section 2.1 does not work, we consider an example of a Gaussian linear model. Specifically, we will show that ordinary least squares estimator in this setting has smaller residuals on training data as opposed to test data.

The setting. Let $X \in \mathbb{R}^{n \times d}$ be a fixed design matrix[1], which will be shared by both training and test data. Let the responses for the training data $Y \in \mathbb{R}^n$ be generated by a Gaussian linear model $Y = X\theta + \epsilon$ where we assume $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Similarly, we assume that the test data is generated $\tilde{Y} = X\theta + \tilde{\epsilon} \in \mathbb{R}^n$ with $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. Here, $\epsilon, \tilde{\epsilon}$ are independent. Thus we study the train and test data using the same X, but with "fresh" noise terms ϵ.

Recall that the we have a closed form solution of the least squares estimator of the parameter $\theta \in \mathbb{R}^d$

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} ||Y - X\theta||^2 = (X^T X)^{-1} X^T Y,$$

and let

$$\hat{Y} = \hat{X}\theta = X(X^T X)^{-1} X^T Y = HY,$$

where the projection $H := X(X^T X)^{-1} X^T$ is the so-called "hat matrix"[2] of the linear regression problem, mapping the training data onto the space of predictions. Furthermore, we introduce the training residuals as follows

$$r = Y - \hat{Y} \in \mathbb{R}^n.$$

Similarly, test residuals are given by

$$\tilde{r} = \tilde{Y} - \hat{Y} \in \mathbb{R}^n.$$

We will show that the magnitude of the test residuals $||\tilde{r}||$ is in general larger than the residuals of the training data $||r||$. We establish bounds in the following proposition:

Proposition 2. *We have*

(i)

$$\mathbb{E}\frac{||r||^2}{n} = \frac{n-d}{n}\sigma^2$$

[1]Note that this does not exactly match the setting in Section 2.1 as X is non-random here and is shared between training and test data. This is done only for convenience, and this detail is in fact immaterial.

[2]The name comes from the fact that it maps Y to the "hat version" $\hat{Y}$.

(ii)

$$\mathbb{E}\frac{||\tilde{r}||^2}{n} = \frac{n+d}{n}\sigma^2$$

(iii)

$$\mathbb{E}\frac{||Y - X\theta||^2}{n} = \sigma^2$$

Observe that, as promised, the quantity in (ii) is larger than quantity in (i), thus, residuals on the training data are optimistic and differ from the residuals on test data. Moreover, note that the θ is the right model for predicting Y and it corresponds to MSE of σ^2 (see (iii)). The quantity in (i) is smaller, which further shows that least squares overfit the training data.

Proof of Proposition 2. We prove each assertion separately.

Proof of (i). By definition of training residuals r, we have

$$\mathbb{E}\,||r||^2 = \mathbb{E}\left|\left|Y - \hat{Y}\right|\right|^2 = \mathbb{E}\,||Y - HY||^2 = \mathbb{E}\,||(I-H)Y||^2$$

Note that in the above, Y is random Gaussian vector, and $(I-H)$ is a fixed matrix. Then, $(I-H)Y$ is distributed as follows

$$(I-H)Y \sim \mathcal{N}\left((I-H)X\theta, (I-H)(I-H)^T\sigma^2\right).$$

From HW2, we know that H is a projection matrix onto the columb space of X. Thus, $HX = X$ and $(I-H)X = 0$, i.e., $(I-H)Y$ is zero-mean. Moreover, $(I-H)$ is thus an orthogonal projection onto the orthocomplement of the column space of X. Thus, $(I-H)(I-H)^\top = (I-H)$ (as projection matrices are idempotent). Thus,

$$(I-H)Y \sim \mathcal{N}\left(0, \sigma^2(I-H)\right).$$

Recall that we are only interested in the expected squared norm $(I-H)Y$. By the rotation invariance of Gaussian vectors, the expected squared norm of $(I-H)Y$ is the same as of a vector sampled from a zero-mean Gaussian with a diagonal covariance that has same eigenvalues as $\sigma^2(I-H)$. As $(I-H)$ is an orthogonal projection onto the orthocomplement of the column space of X, which has dimension $n-d$, it has eigenvalue 1 with multiplicity $n-d$ and eigenvalue 0 with multiplicity d. This implies

$$\mathbb{E}\,||r||^2 = (n-d)\sigma^2.$$

Proof of (ii). By definintion of training residuals $\tilde{r}$, we have

$$\begin{aligned}
\mathbb{E}\,||\tilde{r}||^2 &= \mathbb{E}\,||X\theta + \tilde{\epsilon} - H(X\theta + \epsilon)||^2 \\
&=^{(a)} \mathbb{E}\,||\tilde{\epsilon} - H\epsilon||^2 \\
&=^{(b)} \mathbb{E}\,||\tilde{\epsilon}||^2 + \mathbb{E}\,||H\epsilon||^2,
\end{aligned}$$

where in (a) we used the fact that $HX\theta = X\theta$, since H projects X onto itself, and in (b) we used the fact that $\tilde{\epsilon}$ and ϵ are independent. Note that, in the above

$$\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2 I), \quad \tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2 HH^T).$$

Similarly to the reasoning for part (i), we note that HH^T has eigenvalue 1 with multiplicity d, and eigenvalue 0 with multiplicity $n-d$. Thus,

$$\mathbb{E}\,||H\epsilon||^2 = d\sigma^2,$$

and hence

$$\mathbb{E}\,||\tilde{r}||^2 = n\sigma^2 + d\sigma^2,$$

as desired.

Proof of (iii). By definintion of Guassian linear model, we have

$$\mathbb{E}\,||Y - X\theta||^2 = \mathbb{E}\,||\epsilon||^2 = n\sigma^2.$$

□

An immediate corollary of the proposition above is

Corollary 3 (AIC)**.** *Take*

$$\widehat{MSE} = \left|\left|\hat{Y} - Y\right|\right|^2 + 2d\hat{\sigma}^2,$$

where $\hat{\sigma}^2 = \frac{1}{n-d}\left|\left|\hat{Y} - Y\right|\right|^2$ *(unbiased estimate for* σ^2*). Then,* $\widehat{MSE}$ *is an unbiased estimator for* $\mathbb{E}\left|\left|\tilde{Y} - \hat{Y}\right|\right|^2$.

4 Conformal Prediction

[Only covered high-level introduction in this lecture.] Conformal prediction is a general framework for confidence intervals in prediction problems. In this class, we will primarily focus on the core statistical idea behind the method.

Idea. Consider $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the space of all possible values of y. Intuitively, we want to check if, by setting $Y_{\text{test}} = y$, the datapoint $(X_{\text{test}}, Y_{\text{test}})$ looks "consistent" with the training data. If it does, then we include y in the CI we are constructing; otherwise, $y \notin$ CI.

More formally, we define the family of tests ϕ_y for every $y \in \mathcal{Y}$ of the null hypothesis that $Y_{\text{test}} = y$, and construct the CI-test duality.

$$\text{CI} = \{y : \phi_y(X_{\text{test}}, \mathcal{D}) = 0.\}$$