18.1 Outline
Agenda: Conformal Prediction

1. Full conformal
2. Split conformal
3. Enhanced nonconformity scores
4. Weighted conformal inference

18.2 Full conformal

18.2.1 Prediction intervals
In lecture 8, we talked about the cases in which we use sophisticated machine learning algorithms to make predictions in situations that have tremendous consequences. When the cost of making a wrong prediction is very high, for example being denied college admission, it’s important to understand the reliability of the algorithms we use. A very good way of understanding the reliability or the uncertainty of future prediction is to be able to return prediction intervals.

Suppose we have training data \((X_1, Y_1), \ldots, (X_n, Y_n)\) and a test point \((X_{n+1}, ?)\) (we want to predict the label corresponding to \(X_{n+1}\)). The data are assumed to be exchangeable, e.g. i.i.d. from some distribution \(P_{XY}\). Our goal is to construct marginal distribution free prediction interval \(P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha\) for any (unknown) distribution \(P_{XY}\) and any sample size \(n\). For example, we want to be able to say things like “Based on the candidate’s high school identifier and GPA, SAT scores, and other attributes, the college GPA is predicted in the [3.4,3.8] range”.

We first state the following quantile lemma, which will be useful for our later development.

**Definition 1.** Define the quantile function to be

\[
\text{Quantile}(\beta; F) = \inf\{z : \mathbb{P}\{Z \leq z\} \geq \beta\},
\]

where \(F\) is the cumulative distribution function corresponding to \(\mathbb{P}\).
For a multiset \( v_{1:n} = \{v_1, \ldots, v_n\} \), define

\[
\text{Quantile}(\beta; v_{1:n}) = \text{Quantile}(\beta; \frac{1}{n} \sum_{i=1}^{n} \delta_{v_i}).
\]

**Lemma 1.** If \( V_1, \ldots, V_{n+1} \) are exchangeable, then for any \( \beta \in (0, 1) \),

\[
\mathbb{P}(V_{n+1} \leq \text{Quantile}(\beta; V_{1:n} \cup \{\infty\})) \geq \beta.
\]

If ties between \( V_1, \ldots, V_{n+1} \) occur with probability 0, then above probability is at most \( \beta + 1/(n+1) \).

The key to above lemma is the observation that the rank of \( V_{n+1} \) is uniform over \( \{1, \ldots, n+1\} \) due to the exchangeability of \( V_1, \ldots, V_{n+1} \).

### 18.2.2 Conformal prediction

Suppose we have a training set \( Z_i = (X_i, Y_i), i = 1, \ldots, n \), and a (non-conformity) score function \( S \) with two arguments, a point \((x, y)\) and a multiset \( Z \). A low score of \( S((x, y), Z) \) indicates that \((x, y)\) "conforms" to \( Z \), whereas a high value indicates that \((x, y)\) is atypical relative to the points in \( Z \).

A common example of the score function is given by \( S((x, y), Z) = |y - \hat{\mu}(x)| \), where \( \hat{\mu} : \mathbb{R}^d \rightarrow \mathbb{R} \) is a regression function fitted by running an algorithm \( A \) on \((x, y)\) and \( Z \). We will also assume \( A \) treats its arguments symmetrically.

Now suppose we are given a test point \( x \) and we want to calculate the prediction interval \( \hat{C}_n(x) \) for a potential observation whose covariate has value \( x \).

Define \( V_i^{(x,y)} = S(Z_i, Z_{-i} \cup (x, y)), i = 1, \ldots, n \), and \( V_{n+1}^{(x,y)} = S((x, y), Z_{1:n}) \). Now for a value \( y \), we include \( y \) in \( \hat{C}_n(x) \) if \( V_{n+1}^{(x,y)} \leq \text{Quantile}(1 - \alpha; V_{1:n}^{(x,y)} \cup \{\infty\}) \), where \( V_{1:n}^{(x,y)} = \{V_1^{(x,y)}, \ldots, V_n^{(x,y)}\} \).

The main result is the following:

**Theorem 1.** (Vovk et al. 2005, Lei et al. 2018) Assume that \((X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \ldots, n+1 \) are exchangeable. For any score function \( S \), and any \( \alpha \in (0, 1) \), \( \hat{C}_n \) satisfies

\[
\mathbb{P}(Y_{n+1} \in \hat{C}_n(X_{n+1})) \geq 1 - \alpha.
\]

If ties between \( V_i \)'s occur with probability 0, then above probability is at most \( 1 - \alpha + 1/(n+1) \) (we can also make it equal to \( 1 - \alpha \) by randomization).

This theorem is a simple consequence of the quantile lemma. The only thing to notice is that since \( Z_i \)'s are exchangeable, so are the \( V_i^{(X_{n+1}, Y_{n+1})} \)'s.

### 18.2.3 Computational issues

Overall, the full conformal approach is computationally expensive. For every given value of \( y \), we need to refit the model to decide if we want to include the value \( y \) to our prediction interval. This is the reason why we presented conformal inference differently in lecture 8.
One solution to above issue is to use the split conformal, which we introduce in the next section.

Roughly speaking, the split conformal operates by training $\hat{\mu}$ once and for all on an independent dataset. Our guarantees will be true conditional on $\hat{\mu}$. One drawback of split conformal is that it will be less efficient than the full conformal as it uses only half (or a portion) of the data to find $\hat{\mu}$ and only half (or a portion) of the data is used for calibration.

### 18.3 Split conformal

In split conformal, we compute score function $S$ on an independent proper training set and compute scores on a distinct calibration set, i.e. we get

$$V_i = S(X_i, Y_i), i = 1, \cdots, n,$$  
and

$$V_{n+1} = S(x, y).$$

Then we construct conformal intervals as before.

The following theorem is again a simple consequence of the quantile lemma:

**Theorem 2.** Theorem 1 holds in the setting above if calibration + test points are exchangeable.

Note that although the split conformal is less computationally expensive, full conformal usually produces tighter intervals for in that case we do more with our data (less wasteful).

### 18.4 Better Conformity scores

Now, we are going to concentrate on finding better conformity scores. What we actually want to know are quantiles of the conditional distribution of $Y$ given $x$.

If we would have known that

$$P(q_1(x) \leq Y \leq q_2(x)|X = x) \geq 1 - \alpha,$$


then the band with lower limit $q_1(x)$ (lower quantile) and upper limit $q_2(x)$ (upper quantile) would have been a $(1 - \alpha)th$ prediction interval for $Y$, but as we do not have sufficient data to capture the conditional distribution of $Y|X = x$ for each value of $x$, we have to try a few other methods.

18.4.1 Quantile estimation as a learning task

We can estimate quantiles of the conditional distribution of $Y$ given $X$ by $\hat{f}$ which minimises the following loss function over a class of functions $\mathcal{F}$:

$$\hat{f}(.) = \arg\min_{f \in \mathcal{F}} \sum_i \rho_\alpha(Y_i - f(X_i)) + \mathcal{R}(f),$$

where $\mathcal{R}(.)$ is a possible regularizer and $\rho_\alpha$ is the pinball loss given by

$$\rho_\alpha(y) = y(\alpha - \mathbb{I}(y < 0)).$$

However, the upper and lower quantiles found by the above method would not give valid predictive range for future data points (uncalibrated).

![Test examples](image)

We would need to calibrate the intervals by conformal inference to get a valid coverage.

18.4.2 Conformal Quantile Regression

Yaniv Romano, Evan Patterson and Emmanuel J. Candès [2] proposed the Conformal Quantile Regression which is an improvised method to construct prediction intervals that attains valid coverage in finite samples. As in split conformal prediction, the method begins by splitting the data into a proper training set, indexed by $\mathcal{I}_1$, and a calibration set, indexed by
Given any quantile regression algorithm $A$, we then fit two conditional quantile functions $\hat{q}_{\alpha_{lo}}$ and $\hat{q}_{\alpha_{hi}}$ on the proper training set:

$$(\hat{q}_{\alpha_{lo}}, \hat{q}_{\alpha_{hi}}) \leftarrow A(\{(X_i, Y_i) : i \in I_1\})$$

In the essential next step, we compute conformity scores that quantify the error made by the plug-in prediction interval $\hat{C}(x) = [\hat{q}_{\alpha_{lo}}, \hat{q}_{\alpha_{hi}}]$.

The scores are evaluated on the calibration set as

$$E_i := \max\{\hat{q}_{\alpha_{lo}} - Y_i, Y_i - \hat{q}_{\alpha_{hi}}\}$$

for each $i \in I_2$.

Given new input data $X_{n+1}$, we construct the prediction interval for $Y_{n+1}$ as

$$C(X_{n+1}) = [\hat{q}_{\alpha_{lo}}(X_{n+1}) - Q_{1-\alpha}(E, I_2), \hat{q}_{\alpha_{hi}} + Q_{1-\alpha}(E, I_2)]$$

where $Q_{1-\alpha}(E, I_2)$ is defined to be the $(1-\alpha)(1+1/|I_2|)$-th empirical quantile of $\{E_i : i \in I_2\}$ that conformalizes the plug-in prediction interval.

**Theorem 3.** If $(X_i, Y_i), i = 1, \ldots, n+1$, are exchangeable, then the prediction interval $C(X_{n+1})$ constructed by the split CQR algorithm satisfies

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha.$$

**Proof.** For a proof, follow Theorem 1 of [2].

### 18.4.3 Split Conformal v/s CQR

Recall the Split Conformal method we discussed in lecture 8, where the prediction interval for $Y_{n+1}$ was given by

$$C(X_{n+1}) = [\hat{\mu}(X_{n+1}) - q, \hat{\mu}(X_{n+1}) + q]$$

CQR performs better than Split Conformal in following ways:

- In case of heteroskedasticity, CQR will perform better than split conformal as split conformal produces intervals of constant width however CQR gives adaptive intervals.
We can see that both have similar average coverages (as guaranteed by the theorems we studied), however the average length of CQR is much less than that of split conformal.

- It is intuitive that CQR will provide better conditional coverage than split conformal. As the intervals are not adaptive for split conformal, coverage will be high for some values of $x$ but low for the other values of $x$.

- As $n \to \infty$, prediction ranges provided by CQR are consistent with the quantiles of conditional distribution of $Y$ given $X$. Prediction ranges provided by Split Conformal would be consistent only if $Y = f(X) + \epsilon$ and $\epsilon$ is symmetric.

### 18.4.4 Comparison of Split Conformal and CQR using Medical Services data

Medical Expenditure Panel Survey 2015 comprises of information on 16,000 subjects on 140 features including age, marital status, race, poverty status, functional limitations, health status, and health insurance types, etc. We want to predict health care system utilisation (reflected by the number of visits to doctor’s office and hospital, etc.) using the above covariates.

#### Results on MEPS data

Split conformal and CQR were performed on MEPS data by considering 80% data as training data and remaining 20% data as test data (random splitting). We calculate marginal coverage, conditional coverage (due to Maxime et. al. [3], conditional coverage is measured on the worst slab), and the length of the intervals for both CQR and split conformal method.
We can see that CQR outperforms split conformal by having higher conditional coverage and smaller interval length.

### 18.4.5 Calibration via adaptive coverage

A $1 - \tau$ prediction interval for $y(x)$ is given by

$$[F^{-1}_{y|X}(\tau/2), F^{-1}_{y|X}(1 - \tau/2)].$$

But as the conditional distribution of $Y|X$ is unknown, we give a naive prediction interval by

$$C_{\text{naive}}(x, 1 - \tau) = [\hat{F}^{-1}_{y|X}(\tau/2), \hat{F}^{-1}_{y|X}(1 - \tau/2)],$$

where the estimated conditional CDF $\hat{F}_{y|X}(x)$ may be obtained by fitting several quantile functions under the constraint that $\hat{F}^{-1}_{y|X}(\alpha) \leq \hat{F}^{-1}_{y|X}(\beta) \forall x$ if $\alpha < \beta$ and linearly interpolating in between.

Here, because of the selection bias of $\hat{F}$, the coverage by $C_{\text{naive}}(x, 1 - \tau)$ would not be $1 - \tau$. We instead find $\hat{\tau}$ such that 90% coverage is achieved on the calibration set by the naive prediction interval of level $\hat{\tau}$ and set

$$C(x) = C_{\text{naive}}(x, \hat{\tau}).$$
18.5 Discrete labels

In the next class, we will discuss constructing valid prediction sets in classification problems\[4\]. In a classification problem, we would estimate the conditional probabilities $\hat{\pi}(y|x)$, say by the output of NNet’s softmax layer. As discussed earlier

$$P(Y = i|X = x) \neq \hat{\pi}(i|x)$$

Thus, we need to calibrate the prediction sets.

---

\[4\]In the figure on the top, $\hat{\tau}$ is displaced to be 95% for concreteness. That is just to illustrate the point that $\hat{\tau}$ will be larger than 90%. It may not always be exactly 95% though.
Again we set

\[ C(x) = C^{\text{naive}}(x, \hat{\pi}). \]
Bibliography


