ORIE 4741: Learning with Big Messy Data

Bootstrap and MLE

Professor Udell
Operations Research and Information Engineering
Cornell

October 29, 2017
Estimate sensitivity of prediction

- suppose each \((x_i, y_i) \sim P, \ i = 1, \ldots, n, \ iid\)
- given \(\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}\)
- estimate model \(g_\mathcal{D} : \mathcal{X} \to \mathcal{Y}\)
- use it to make prediction \(g_\mathcal{D}(x)\) for new input \(x\)

**Q:** How sensitive is the prediction to the data set \(\mathcal{D}\)?

**Q:** Can we compute a **confidence interval** for the prediction?
Ideal confidence intervals

for \( k = 1, \ldots \)

- sample new \((x_i^k, y_i^k) \sim P, i = 1, \ldots, n, \) iid to form dataset \( \mathcal{D}_k \)
- estimate model \( g_{\mathcal{D}_k} : \mathcal{X} \rightarrow \mathcal{Y} \)
- use it to make prediction \( g_{\mathcal{D}_k}(x) \) for new input \( x \)

Q: How sensitive is the prediction to the data set \( \mathcal{D} \)?
Ideal confidence intervals

for $k = 1, \ldots$

- sample new $(x_i^k, y_i^k) \sim P$, $i = 1, \ldots, n$, iid to form dataset $D_k$
- estimate model $g_{D_k} : \mathcal{X} \rightarrow \mathcal{Y}$
- use it to make prediction $g_{D_k}(x)$ for new input $x$

Q: How sensitive is the prediction to the data set $D$?
A: Look at histogram of $\{g_{D_k}(x)\}_k$
Ideal confidence intervals

for \( k = 1, \ldots \)

- sample new \((x_i^k, y_i^k) \sim P, \ i = 1, \ldots, n, \ \text{iid}\) to form dataset \( \mathcal{D}_k \)
- estimate model \( g_{\mathcal{D}_k} : \mathcal{X} \rightarrow \mathcal{Y} \)
- use it to make prediction \( g_{\mathcal{D}_k}(x) \) for new input \( x \)

**Q:** How sensitive is the prediction to the data set \( \mathcal{D} \)?

**A:** Look at histogram of \( \{g_{\mathcal{D}_k}(x)\}_k \)

**Q:** Can we compute a confidence interval for the prediction?
Ideal confidence intervals

for $k = 1, \ldots$

- sample new $(x_i^k, y_i^k) \sim P$, $i = 1, \ldots, n$, iid
to form dataset $\mathcal{D}_k$
- estimate model $g_{\mathcal{D}_k} : \mathcal{X} \rightarrow \mathcal{Y}$
- use it to make prediction $g_{\mathcal{D}_k}(x)$ for new input $x$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?
A: Look at histogram of $\{g_{\mathcal{D}_k}(x)\}_k$

Q: Can we compute a confidence interval for the prediction?
A: Look at 95% confidence bound for $\{g_{\mathcal{D}_k}(x)\}_k$
Bootstrap: confidence with limited data

given dataset \( D \), for \( k = 1, \ldots \)

- sample \((x_i^k, y_i^k)\) with replacement from \( D \), \( i = 1, \ldots, n \), to form dataset \( D_k \)
- estimate model \( g_{D_k} : \mathcal{X} \rightarrow \mathcal{Y} \)
- use it to make prediction \( g_{D_k}(x) \) for new input \( x \)

**Q:** How sensitive is the prediction to the data set \( D \)?

**A:** Look at histogram of \( \{g_{D_k}(x)\}_k \)

**Q:** Can we compute a **confidence interval** for the prediction?

**A:** Look at 95\% confidence bound for \( \{g_{D_k}(x)\}_k \)
Bootstrap estimator for the variance

pick a function $h : \mathcal{D} \rightarrow \mathbb{R}$.
we want to estimate how much $h$ varies when applied to finite data sets from the same distribution.

- resample $\mathcal{D}_1, \ldots, \mathcal{D}_K$ from $\mathcal{D}$
- compute $h(\mathcal{D}_1), \ldots, h(\mathcal{D}_K)$
- estimate the mean $\hat{\mu}_h = \frac{1}{K} \sum_{k=1}^{K} h(\mathcal{D}_k)$
- estimate the variance

$$\hat{\sigma}_h = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (h(\mathcal{D}_k) - \hat{\mu}_h)^2}$$
Why does bootstrap work?

sample \((x_i^k, y_i^k)\) with replacement from \(\mathcal{D}\)

\[
\mathbb{P} \left( \left( x_1^1, y_1^1 \right) = (x, y) \right)
\]

\[= \sum_{i=1}^{n} \mathbb{P} \left( \text{picked} \ (x_i, y_i) \ \text{from} \ \mathcal{D} \ \text{and was equal to} \ (x, y) \right) \]

\[= \sum_{i=1}^{n} \mathbb{P} \left( \text{picked} \ (x_i, y_i) \ \text{from} \ \mathcal{D} \right) \mathbb{P} \left( \left( x_i, y_i \right) = (x, y) \right) \]

\[= \sum_{i=1}^{n} \frac{1}{n} \mathbb{P} \left( x, y \right) \]

\[= \frac{1}{n} \mathbb{P} \left( x, y \right) \]

\[= n \frac{1}{n} \mathbb{P} \left( x, y \right) \]

\[= \mathbb{P} \left( x, y \right) \]
Why does bootstrap work?

\( \mathcal{D}_k \) each have the same distribution as \( \mathcal{D} \). So for any function \( h : \mathcal{D} \to \mathbb{R} \),

\[
\mathbb{E}_{\mathcal{D}} \frac{1}{K} \sum_{k=1}^{K} h(\mathcal{D}_k) = \mathbb{E}_{\mathcal{D}} h(\mathcal{D})
\]