ORIE 4741: Learning with Big Messy Data

The Bootstrap and the Bias Variance Tradeoff

Professor Udell

Operations Research and Information Engineering
Cornell

October 16, 2021

Announcements 10/14/21

- section (only yesterday) this week: advanced scikit-learn
- hw3 due this Friday 11:59pm
- hw4 out this weekend, due in two weeks
  - save slip days for emergencies
- begin work on project midterm report

Outline

Bootstrap

Bias variance tradeoff

Why regularization helps

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 $g_{\mathcal{D}_k} : \mathcal{X} \to \mathcal{Y}$
- 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 $\mathcal{D}_k$
- estimate model $g_{\mathcal{D}_k} : \mathcal{X} \to \mathcal{Y}$
- 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 $\{g_{\mathcal{D}_k}(x)\}_k$

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 $g_{\mathcal{D}_k} : \mathcal{X} \to \mathcal{Y}$
- 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 $g_{\mathcal{D}_k} : \mathcal{X} \to \mathcal{Y}$
- 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 $\mathcal{D}$, for $k = 1, \dots$

- sample (x_i^k, y_i^k) **with replacement** from $\mathcal{D}$, $i = 1, \dots, n$, to form dataset $\mathcal{D}_k$
- estimate model $g_{\mathcal{D}_k} : \mathcal{X} \to \mathcal{Y}$
- 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 estimator for the variance

pick a function $h : \mathcal{D} \to \mathbf{R}$.
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

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

Demo: The bootstrap

`https://github.com/ORIE4741/demos/bootstrap.ipynb`

Why does bootstrap work?

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

$$
\begin{aligned}
& \mathbb{P}\left((x_1^1, y_1^1) = (x, y)\right) \\
= \ & \sum_{i=1}^{n} \mathbb{P}(\text{picked } (x_i, y_i) \text{ from } \mathcal{D} \text{ and was equal to } (x, y)) \\
= \ & \sum_{i=1}^{n} \mathbb{P}(\text{picked } (x_i, y_i) \text{ from } \mathcal{D})\, \mathbb{P}((x_i, y_i) = (x, y)) \\
= \ & \sum_{i=1}^{n} \frac{1}{n}\, \mathbb{P}(x, y) \\
= \ & n \frac{1}{n}\, \mathbb{P}(x, y) \\
= \ & \mathbb{P}(x, y)
\end{aligned}
$$

Why does bootstrap work?

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

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

References

- The Bootstrap: `http://www.stat.cmu.edu/~larry/=stat705/Lecture13.pdf`. Wasserman, CMU Stat 705.

Outline

Bias variance tradeoff

analyze out of sample square error:

$$E_{\text{out}}(g_{\mathcal{D}}) = \mathbb{E}_{(x,y)\sim P}(y - g_{\mathcal{D}}(x))^2$$

take expectation over all data sets $\mathcal{D}$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} E_{\text{out}}(g_{\mathcal{D}}) &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(x,y)\sim P}(y - g_{\mathcal{D}}(x))^2\right] \\
&= \mathbb{E}_{(x,y)\sim P}\left[\mathbb{E}_{\mathcal{D}}(y - g_{\mathcal{D}}(x))^2\right] \\
&= \mathbb{E}_{(x,y)\sim P}\left[\mathbb{E}_{\mathcal{D}}\left[(g_{\mathcal{D}}(x))^2\right] - 2y\mathbb{E}_{\mathcal{D}}\left[g_{\mathcal{D}}(x)\right] + y^2\right]
\end{aligned}$$

Bias variance tradeoff: average function

define the **average function** $\bar{g}(x) = \mathbb{E}_{\mathcal{D}}[g_{\mathcal{D}}(x)]$

- depends on test point x
- independent of the data set $\mathcal{D}$ used to choose the model g

the average function is a **conceptual** tool, not a computational tool

could (theoretically) estimate the average function by

- generating many data sets $\mathcal{D}_1, \ldots, \mathcal{D}_K$
- fitting a model g_i to each data set $\mathcal{D}_i$, $i = 1, \ldots, K$
- computing $\bar{g}(x) = \frac{1}{K} \sum_{i=1}^{K} g_i(x)$

Bias variance tradeoff: average function

define the **average function** $\bar{g}(x) = \mathbb{E}_{\mathcal{D}}[g_{\mathcal{D}}(x)]$

- depends on test point x
- independent of the data set $\mathcal{D}$ used to choose the model g

the average function is a **conceptual** tool, not a computational tool

could (theoretically) estimate the average function by

- generating many data sets $\mathcal{D}_1, \ldots, \mathcal{D}_K$
- fitting a model g_i to each data set $\mathcal{D}_i$, $i = 1, \ldots, K$
- computing $\bar{g}(x) = \frac{1}{K}\sum_{i=1}^{K} g_i(x)$

Q: is the average model $\bar{g}$ always in the hypothesis set $\mathcal{H}$?

A. yes

B. no

Bias variance tradeoff

use average function to rewrite out of sample error:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} E_{\text{out}}(g_{\mathcal{D}}) &= \mathbb{E}_{(x,y)\sim P}\left[\mathbb{E}_{\mathcal{D}}\left[g_{\mathcal{D}}(x)^2\right] - 2y\bar{g}(x) + y^2\right] \\
&= \mathbb{E}_{(x,y)\sim P}\left[\mathbb{E}_{\mathcal{D}}\left[g_{\mathcal{D}}(x)^2\right] - \bar{g}(x)^2 \right. \\
&\qquad \left. +\bar{g}(x)^2 - 2y\bar{g}(x) + y^2\right] \\
&= \mathbb{E}_{(x,y)\sim P}\left[\mathbb{E}_{\mathcal{D}}\left[(g_{\mathcal{D}}(x) - \bar{g}(x))^2\right] + (\bar{g}(x) - y)^2\right]
\end{aligned}$$

($\bar{g}(x)$ is constant wrt $\mathcal{D}$)

Bias variance tradeoff

now suppose $y = f(x) + \varepsilon$ where the noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is iid and independent of x.

$$\begin{aligned}
\mathbb{E}_{(x,y)}[(\bar{g}(x) - y)^2] &= \mathbb{E}_{(x,\varepsilon)}[(\bar{g}(x) - f(x) - \varepsilon)^2] \\
&= \mathbb{E}_{(x,\varepsilon)}[(\bar{g}(x) - f(x))^2 + 2\varepsilon(\bar{g}(x) - f(x)) + \varepsilon^2] \\
&= \mathbb{E}_x[(\bar{g}(x) - f(x))^2] + \sigma^2
\end{aligned}$$

so

$$\mathbb{E}_{\mathcal{D}} E_{\text{out}}(g_{\mathcal{D}}) = \mathbb{E}_x \left[\underbrace{\mathbb{E}_{\mathcal{D}} \left[(g_{\mathcal{D}}(x) - \bar{g}(x))^2 \right]}_{\textbf{var(x)}} + \underbrace{(\bar{g}(x) - f(x))^2}_{\textbf{bias}^2\textbf{(x)}} \right] + \underbrace{\sigma^2}_{\textbf{noise}}$$

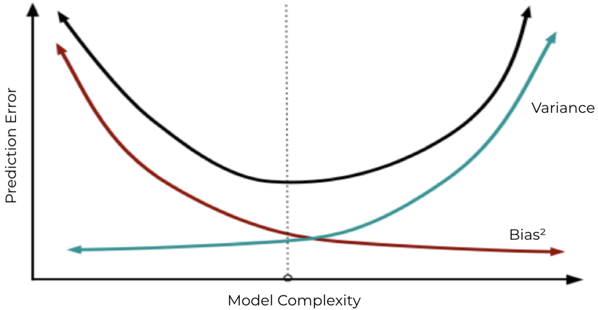
and

$$\mathbb{E}_{\mathcal{D}} E_{\text{out}}(g_{\mathcal{D}}) = \mathbb{E}_x \left[\textbf{bias}^2\textbf{(x)} + \textbf{var(x)} \right] + \textbf{noise} = \textbf{bias}^2 + \textbf{var} + \textbf{noise}$$

Bias variance tradeoff

$$\mathbb{E}_{\mathcal{D}} E_{\text{out}}(g_{\mathcal{D}}) = \mathbb{E}_{(x,y)\sim P}\left[\underbrace{\mathbb{E}_{\mathcal{D}}\left[(g_{\mathcal{D}}(x) - \bar{g}(x))^2\right]}_{\textbf{var(x)}} + \underbrace{(\bar{g}(x) - y)^2}_{\textbf{bias}^2\textbf{(x)}}\right]$$

- we want flexible, responsive models to reduce **bias**
- we want rigid, constrained models to reduce **var**



Outline

Bias variance tradeoff for regression

- suppose $y = Xw^\natural + \epsilon$
- $X = U\Sigma V^T$ is the SVD of X

$$w^{\text{ridge}} = \sum_{i=1}^{d} v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^T y, \qquad w^{\text{lsq}} = \sum_{i=1}^{d} v_i \frac{1}{\sigma_i} u_i^T y$$

Bias variance tradeoff: least squares regresion

- suppose $y = Xw^\natural + \varepsilon$, $\varepsilon_i \sim \mathcal{N}(0,1)$ iid for $i = 1, \dots, n$
- different samples of datasets $\mathcal{D}$ have same X, different ε
- $X = U\Sigma V^T$ is the SVD of X

- true model
$$f(x) = x^T w^\natural$$
- predictions based on data $\mathcal{D}$:
$$\begin{aligned} g_{\mathcal{D}}(x) &= x^T(X^TX)^{-1}X^Ty = x^T(X^TX)^{-1}X^T(Xw^\natural + \varepsilon) \\ &= x^Tw^\natural + x^T(X^TX)^{-1}X^T\varepsilon \end{aligned}$$
- expectation of predictions over random data:
$$\bar{g}(x) = \mathbb{E}_{\mathcal{D}}\left[g_{\mathcal{D}}(x)\right] = x^T w^\natural$$

Bias variance tradeoff: least squares regresion

so

$$
\begin{aligned}
\textbf{bias}(\mathbf{x}) &= f(x) - \bar{g}(x) = 0 \\
\textbf{var}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}}\left[(g_{\mathcal{D}}(x) - \bar{g}(x))^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[x^T(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}x\right] \\
&= x^T(X^TX)^{-1}X^T\mathbb{E}_{\mathcal{D}}\left[\varepsilon\varepsilon^T\right]X(X^TX)^{-1}x \\
&= x^T(X^TX)^{-1}X^TIX(X^TX)^{-1}x \\
&= x^T(X^TX)^{-1}X^TX(X^TX)^{-1}x \\
&= x^T(X^TX)^{-1}x \\
&= x^T\left(\sum_{i=1}^{d} v_i\frac{1}{\sigma_i^2}v_i^T\right)x
\end{aligned}
$$

Bias variance tradeoff: ridge regresion

- suppose $y = Xw^\natural + \varepsilon$, $\varepsilon_i \sim \mathcal{N}(0,1)$ iid for $i = 1, \ldots, n$
- different samples of datasets $\mathcal{D}$ have same X, different ε
- $X = U\Sigma V^T$ is the SVD of X

$$
\begin{aligned}
f(x) &= x^T w^\natural \\
g_{\mathcal{D}}(x) &= x^T w^{\text{ridge}} = x^T (X^T X + \lambda I)^{-1} X^T y \\
&= x^T (X^T X + \lambda I)^{-1} X^T (Xw^\natural + \varepsilon) \\
\bar{g}(x) &= \mathbb{E}_{\mathcal{D}}\left[g_{\mathcal{D}}(x)\right] = x^T (X^T X + \lambda I)^{-1} X^T X w^\natural
\end{aligned}
$$

Bias variance tradeoff: ridge regresion

so

$$\begin{aligned}
\textbf{bias}(\mathbf{x}) &= f(x) - \bar{g}(x) = x^T((X^TX + \lambda I)^{-1}X^TX - I)w^\natural \\
\textbf{var}(\mathbf{x}) &= \mathbb{E}_\mathcal{D}\left[(g_\mathcal{D}(x) - \bar{g}(x))^2\right] \\
&= \mathbb{E}_\mathcal{D}\left[x^T(X^TX + \lambda I)^{-1}X^T\varepsilon\varepsilon^TX(X^TX + \lambda I)^{-1}x\right] \\
&= x^T(X^TX + \lambda I)^{-1}X^T\mathbb{E}_\mathcal{D}\left[\varepsilon\varepsilon^T\right]X(X^TX + \lambda I)^{-1}x \\
&= x^T(X^TX + \lambda I)^{-1}X^TIX(X^TX + \lambda I)^{-1}x \\
&= x^T(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}x \\
&= x^T\left(\sum_{i=1}^d v_i \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2} v_i^T\right)x
\end{aligned}$$