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

Q: How sensitive is the prediction to the data set $D$?
for $k = 1, \ldots$

- sample new $(x^k_i, y^k_i) \sim P$, $i = 1, \ldots, n$, iid to form new $D_k$
- estimate model $g_{D_k} : \mathcal{X} \to \mathcal{Y}$
- 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?
Ideal confidence intervals

for $k = 1, \ldots$

- sample new $(x_i^k, y_i^k) \sim P$, $i = 1, \ldots, n$, iid to form new $D_k$
- 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: confidence with limited data

for $k = 1, \ldots$

- sample $(x_i^k, y_i^k)$ with replacement from $\mathcal{D}$, $i = 1, \ldots, n$, to form $\mathcal{D}_k$
- 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 $\mathcal{D}$?
Bootstrap: confidence with limited data

for \( k = 1, \ldots \)

- sample \( (x_i^k, y_i^k) \) with replacement from \( D, i = 1, \ldots, n \), to form \( D_k \)
- 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?
Bootstrap: confidence with limited data

for \( k = 1, \ldots \)

- sample \((x_i^k, y_i^k)\) with replacement from \( D, i = 1, \ldots, n \), to form \( D_k \)
- 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 \( h : D \rightarrow \mathbb{R} \).
we want to estimate how much \( h \) varies when applied to finite data sets from the same distribution.

- resample \( D_1, \ldots, D_K \) from \( D \)
- compute \( h(D_1), \ldots, h(D_K) \)
- estimate the mean \( \hat{\mu}_h = \frac{1}{K} \sum_{k=1}^{K} h(D_k) \)
- estimate the variance

\[
\hat{\sigma}_h = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (h(D_k) - \hat{\mu}_h)^2}
\]
Demo: Bootstrap

https://github.com/ORIE4741/demos/linear.ipynb
Why does bootstrap work?

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

\[
\mathbb{P} \left( (x_1^1, y_1^1) = (x, y) \right) = \sum_{i=1}^{n} \mathbb{P}(\text{picked } (x_i, y_i) \text{ from } \mathcal{D} \text{ and it was equal to } (x, y))
\]

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

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

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

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

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

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

Probabilistic setup

- suppose you know a function \( p : \mathbb{R} \rightarrow [0, 1] \) so that
  \[
P(y_i = y \mid x_i, w) = p(y; x_i, w)
  \]
- for example, if \( y_i = w^T x_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \), then
  \[
P(y_i = y \mid x_i, w) \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right)
  \]
- likelihood of data given parameter \( w \) is
  \[
  L(D; w) = \prod_{i=1}^{n} P(y_i = y \mid x_i, w)
  \]
- for example, for linear model with Gaussian error,
  \[
  L(D; w) \sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right)
  \]
Maximum Likelihood Estimation (MLE)

**MLE:** choose \( w \) to maximize \( L(\mathcal{D}; w) \)

- **likelihood**
  \[
  L(\mathcal{D}; w) = \prod_{i=1}^{n} p(y_i; x_i, w)
  \]

- **negative log likelihood**
  \[
  \ell(\mathcal{D}; w) = -\log L(\mathcal{D}; w)
  \]

- maximize \( L(\mathcal{D}; w) \) \iff\ maximize \( \ell(\mathcal{D}; w) \)
Example: Maximum Likelihood Estimation (MLE)

- for linear model with Gaussian error,

\[
\ell(D; w) \sim - \log \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right)
\]

\[
= \sum_{i=1}^{n} - \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right)
\]

\[
= \sum_{i=1}^{n} \left( \frac{1}{2} \log(2\pi\sigma^2) - \log \left( \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right) \right)
\]

\[
= \frac{n}{2} \log(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{1}{2\sigma^2} (y_i - w^T x_i)^2
\]

\[
= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2
\]

- so maximize \( L(D; w) \iff \) minimize \( \sum_{i=1}^{n} (y_i - w^T x_i)^2 \)
what if I have beliefs about what \( w \) should be before I begin?

- \( w \) should be small
- \( w \) should be sparse
- \( w \) should be nonnegative

**idea:** impose **prior** on \( w \) to specify

\[ \mathbb{P}(w) \]

before seeing any data
Maximum-a-posteriori estimation

after I see data, compute posterior probability

\[ P(D; w) = P(D | w) P(w) \]

maximum a posteriori (MAP estimation): choose \( w \) to maximize posterior probability
Maximum-a-posteriori estimation

after I see data, compute posterior probability

\[ P(D; w) = P(D \mid w) P(w) \]

maximum a posteriori (MAP estimation): choose \( w \) to maximize posterior probability

n.b. this is not what a true Bayesian would do

(see, e.g., Bishop, Pattern Recognition and Machine Learning)
Ridge regression: interpretation as MAP estimator

- prior probability of model \( w \sim \mathcal{N}(0, I_d) \)
- noise \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \), \( i = 1, \ldots, n \)
- response \( y_i = w^T x_i + \epsilon_i \), \( i = 1, \ldots, n \)

\[
\mathbb{P}(D; w) = \mathbb{P}(D \mid w) \mathbb{P}(w)
\]

\[
\approx \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( \frac{-(y_i - w^T x_i)^2}{2\sigma^2} \right) \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp \left( \frac{-w_i^2}{2} \right)
\]

\[
= (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \left( \exp \left( \frac{-(y_i - w^T x_i)^2}{2\sigma^2} \right) \right) (2\pi)^{-\frac{d}{2}} \prod_{i=1}^{d} \left( \exp \left( \frac{-w_i^2}{2} \right) \right)
\]

\[
\ell(D; w) = -\log \left( \mathbb{P}(D; w) \right)
\]

\[
= \frac{n}{2} \log(2\pi\sigma^2) + \frac{d}{2} \log(2\pi) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \frac{1}{2} \sum_{i=1}^{d} w_i^2
\]
Ridge regression: interpretation as MAP estimator

- prior probability of model $w \sim \mathcal{N}(0, I_d)$
- noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$
- response $y_i = w^T x_i + \epsilon_i, i = 1, \ldots, n$

\[
\mathbb{P}(D; w) = \mathbb{P}(D \mid w) \mathbb{P}(w)
\]
\[
\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w_i^2}{2}\right)
\]
\[
= (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \left(\exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)\right) (2\pi)^{-\frac{d}{2}} \prod_{i=1}^{d} \left(\exp\left(-\frac{w_i^2}{2}\right)\right)
\]
\[
\ell(D; w) = -\log(\mathbb{P}(D; w))
\]
\[
= \frac{n}{2} \log(2\pi\sigma^2) + \frac{d}{2} \log(2\pi) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \frac{1}{2} \sum_{i=1}^{d} w_i^2
\]

...aha! and we have ridge regression with $\lambda = \sigma^2$