Estimation:
Sample Averages, Bias, and Concentration Inequalities

CMPUT 267: Basics of Machine Learning

Logistics

Outline

1. Recap
2. Sample Complexity

Recap

- The **variance** $\mathrm{Var}[X]$ of a random variable X is its expected squared distance from the mean
- An **estimator** is a random variable representing a procedure for estimating the value of an unobserved quantity based on observed data
- **Concentration inequalities** let us bound the probability of a given estimator being at least ϵ from the estimated quantity
- An estimator is **consistent** if it **converges in probability** to the estimated quantity

Confidence intervals

- We want to obtain a confidence interval around our estimate - we want the difference from the expected value to be small, and be consistently small.
- We would like to be able to claim $\mathbf{Pr}\left(\left|\bar{X} - \mu\right| < \epsilon\right) > 1 - \delta$ for some $\delta, \epsilon > 0$
- This tells us that $\mathbb{E}[\bar{X}] \in \{\bar{X} - \epsilon, \bar{X} + \epsilon\}$ with a large probability, $1 - \delta$
- Confidence level: δ, width of interval: ϵ

Hoeffding's Inequality

Theorem: Hoeffding's Inequality

Suppose that $X_1, \ldots, X_n$ are distributed i.i.d, with $a \leq X_i \leq b$.

Then for any $\epsilon > 0$,

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Equivalently, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}\right) \geq 1 - \delta.$

Chebyshev's Inequality

Theorem: Chebyshev's Inequality

Suppose that $X_1, \ldots, X_n$ are distributed i.i.d. with variance σ^2. Then for any $\epsilon > 0$,

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Equivalently, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{\sigma^2}{\delta n}}\right) \geq 1 - \delta.$

When to Use Chebyshev, When to Use Hoeffding?

- **Popoviciu's inequality:** If $a \leq X_i \leq b$, then $\text{Var}[X_i] \leq \frac{1}{4}(b-a)^2$
- Hoeffding's inequality gives $\epsilon = (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}} = \sqrt{\frac{\ln(2/\delta)}{2}}(b-a)\sqrt{\frac{1}{n}}$;

  Chebyshev's inequality gives $\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} \leq \sqrt{\frac{(b-a)^2}{4\delta n}} = \frac{1}{2\sqrt{\delta}}(b-a)\sqrt{\frac{1}{n}}$
- **Hoeffding's inequality** gives a **tighter bound***, but it can only be used on **bounded** random variables

  * whenever $\sqrt{\frac{\ln(2/\delta)}{2}} < \frac{1}{2\sqrt{\delta}} \iff \delta < \sim 0.232$
- **Chebyshev's inequality** can be applied even for **unbounded** variables

Consistency

Definition: A sequence of random variables X_n **converges in probability** to a random variable X (written $X_n \xrightarrow{p} X$) if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0.$$

Definition: An estimator $\hat{X}$ for a quantity X is **consistent** if $\hat{X} \xrightarrow{p} X$.

Convergence Rate via Chebyshev

The **convergence rate** indicates how quickly the error in an estimator decays as the number of samples grows.

Example: Estimating mean of a distribution using $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Recall that **Chebyshev's inequality** guarantees

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{\sigma^2}{\delta n}}\right) \geq 1 - \delta$$

- Convergence rate is thus $O\left(1/\sqrt{n}\right)$

Sample Complexity

Definition:
The sample complexity of an estimator is the number of samples required to guarantee an expected error of at most ϵ with probability $1 - \delta$, for given δ and ϵ.

Chebyshev gives

$$\epsilon = \sqrt{\frac{\sigma^2}{\delta n}}$$

$$\iff \sqrt{n} = \frac{\sigma}{\epsilon\sqrt{\delta}}$$

$$\iff n = \frac{\sigma^2}{\epsilon^2\delta}$$

Sample Complexity

Definition:
The **sample complexity** of an estimator is the number of samples required to guarantee an expected error of at most ϵ with probability $1-\delta$, for given δ and ϵ.

For $\delta = 0.05$, **Chebyshev** gives

$$\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} = \frac{1}{\sqrt{0.05}}\frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \epsilon = 4.47\frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \sqrt{n} = 4.47\frac{\sigma}{\epsilon}$$

$$\iff \quad n = 19.98\frac{\sigma^2}{\epsilon^2}$$

Sample Complexity

Definition:
The **sample complexity** of an estimator is the number of samples required to guarantee an expected error of at most ϵ with probability $1 - \delta$, for given δ and ϵ.

For $\delta = 0.05$, **Chebyshev** gives

$$\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} = \frac{1}{\sqrt{0.05}} \frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \epsilon = 4.47 \frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \sqrt{n} = 4.47 \frac{\sigma}{\epsilon}$$

$$\iff \quad n = 19.98 \frac{\sigma^2}{\epsilon^2}$$

With **Gaussian assumption** and $\delta = 0.05$,

$$\epsilon = 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \sqrt{n} = 1.96 \frac{\sigma}{\epsilon}$$

$$\iff \quad n = 3.84 \frac{\sigma^2}{\epsilon^2}$$

How good is an estimator?

- **Bias:** whether an estimator is correct **in expectation**
- **Consistency:** whether an estimator is correct **in the limit of infinite data**
- **Convergence rate:** how fast the estimator **approaches its own mean**
  - For an **unbiased** estimator, this is also how fast its **error bounds** shrink
- We don't necessarily care about an estimator's being unbiased.
  - Often, what we care about is our estimator's **accuracy in expectation**

Mean-Squared Error

- We don't necessarily care about an estimator's being unbiased.
  - Often, what we care about is our estimator's **accuracy in expectation**

Definition: **Mean squared error** of an estimator $\hat{X}$ of a quantity X:

$$\mathrm{MSE}(\hat{X}) = \mathbb{E}\left[(\hat{X} - \mathbb{E}[X])^2\right]$$

different!

Bias-Variance Decomposition

Sometimes a biased estimator can be closer to the estimated quantity than an unbiased one.

$$
\begin{aligned}
MSE(\hat{X}) &= \mathbb{E}[(\hat{X} - \mathbb{E}[X])^2] = \mathbb{E}[(\hat{X} - \mu)^2] && \mu = \mathbb{E}[X] \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}] + \mathbb{E}[\hat{X}] - \mu)^2] && -\mathbb{E}[\hat{X}] + \mathbb{E}[\hat{X}] = 0 \\
&= \mathbb{E}[((\hat{X} - \mathbb{E}[\hat{X}]) + b)^2] && b = \mathrm{Bias}(\hat{X}) = \mathbb{E}[\hat{X}] - \mu \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])^2 + 2b(\hat{X} - \mathbb{E}[\hat{X}]) + b^2] \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])^2] + \mathbb{E}[2b(\hat{X} - \mathbb{E}[\hat{X}])] + \mathbb{E}[b^2] && \text{linearity of } \mathbb{E} \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])^2] + 2b\mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])] + b^2 && \text{constants come out of } \mathbb{E} \\
&= \mathrm{Var}[\hat{X}] + 2b\mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])] + b^2 && \text{def. variance} \\
&= \mathrm{Var}[\hat{X}] + 2b(\mathbb{E}[\hat{X}] - \mathbb{E}[\hat{X}]) + b^2 && \text{linearity of } \mathbb{E} \\
&= \mathrm{Var}[\hat{X}] + b^2 \\
&= \mathrm{Var}[\hat{X}] + \mathrm{Bias}(\hat{X})^2 \qquad \blacksquare
\end{aligned}
$$

Bias-Variance Tradeoff

$$\mathrm{MSE}(\hat{X}) = \mathrm{Var}[\hat{X}] + \mathrm{Bias}(\hat{X})^2$$

- If we can decrease bias without increasing variance, error goes down
- If we can decrease variance without increasing bias, error goes down
- **Question:** Would we ever want to **increase bias**?
- *YES.* If we can increase (squared) bias in a way that **decreases variance more**, then error goes down!
  - **Interpretation:** Biasing the estimator toward values that are **more likely to be true** (based on **prior information)**

Downward-biased Mean Estimation

Example: Let's estimate μ given i.i.d $X_1, \ldots, X_n$ with $\mathbb{E}[X_i] = \mu$ using: $Y = \frac{1}{n+100} \sum_{i=1}^{n} X_i$

This estimator is **biased**:

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{1}{n+100}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n+100}\sum_{i=1}^{n} \mathbb{E}[X_i]$$

$$= \frac{n}{n+100}\mu$$

$$\mathrm{Bias}(Y) = \frac{n}{n+100}\mu - \mu = \frac{-100}{n+100}\mu$$

This estimator has **low variance**:

$$\mathrm{Var}(Y) = \mathrm{Var}\left[\frac{1}{n+100}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{(n+100)^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{(n+100)^2}\sum_{i=1}^{n} \mathrm{Var}[X_i]$$

$$= \frac{n}{(n+100)^2}\sigma^2$$

Estimating μ Near 0

Example: Suppose that $\sigma = 1$, $n = 10$, and $\mu = 0.1$

$\mathrm{Bias}(\bar{X}) = 0$

$$\mathrm{MSE}(\bar{X}) = \mathrm{Var}(\bar{X}) + \mathrm{Bias}(\bar{X})^2$$

$$= \mathrm{Var}(\bar{X})$$

$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$

$$= \frac{1}{10}$$

$$\mathrm{MSE}(Y) = \mathrm{Var}(Y) + \mathrm{Bias}(Y)^2$$

$$= \frac{n}{(n+100)^2}\sigma^2 + \left(\frac{100}{n+100}\mu\right)^2$$

$$= \frac{10}{110^2} + \left(\frac{100}{110}0.1\right)^2$$

$$\approx 9 \times 10^{-4}$$

Estimating μ

Example: Suppose that $\sigma = 1$, $n = 10$, and $\mu = 5$

$\text{Bias}(\bar{X}) = 0$

$$\begin{aligned}\text{MSE}(\bar{X}) &= \text{Var}(\bar{X}) + \text{Bias}(\bar{X})^2 \\ &= \text{Var}(\bar{X}) \\ &= \frac{1}{10}\end{aligned}$$

$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

$$\begin{aligned}\text{MSE}(Y) &= \text{Var}(Y) + \text{Bias}(Y)^2 \\ &= \frac{n}{(n+100)^2}\sigma^2 + \left(\frac{100}{n+100}\mu\right)^2 \\ &= \frac{10}{110^2} + \left(\frac{100}{110}5\right)^2 \\ &\approx 20.7\end{aligned}$$