Probability Bounds

- probability bounds
- deviations from the mean
- Markov/Chebyshev
- law of large numbers
- central limit theorem
- Hoeffding

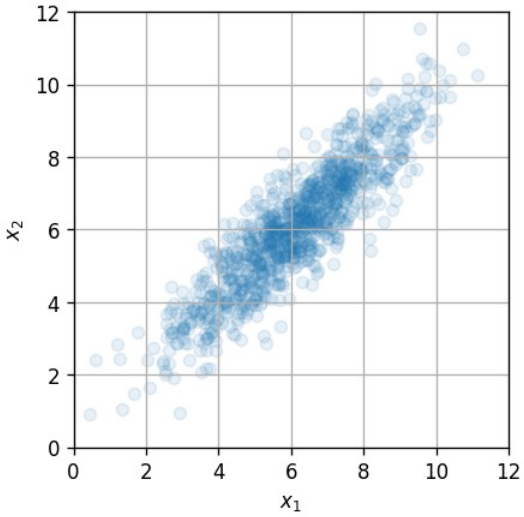
Why Probability Bounds in ML?

- *supervised* machine learning is about **learning functions from data**

  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ • learn $p(\boldsymbol{x}, y)$

- *unsupervised* machine learning is about finding interesting patterns in data

  $\mathcal{D} = \{(x_i)\}_{i=1}^n$ • learn $p(\boldsymbol{x})$



- how well can we approximate $p(\boldsymbol{x})$ or $p(\boldsymbol{x}, y)$?
  - intuition: more samples, better job learning $p(\boldsymbol{x})$.
  - simpler question - how well can we estimate $E[\boldsymbol{x}]$?

- why bounds?
  - always want to know how well we are doing
  - exact expressions are hard/impossible to derive, and don't provide insight
  - bounds allow for more general assumptions

Upper Bounds, Lower Bounds, Sample Complexity

- upper bounds vs. lower bounds

$$f_l(x) \leq f(x) \leq f_u(x)$$

$$n^n e^{-n} \leq n! \leq n^n$$

- sample complexity

the number of samples required by a learning algorithm to achieve error/loss ϵ with probability $1 - \delta$.

- typical machine learning bound:

Theorem 1: Alg. 1 achieves loss less than ϵ with with probability greater than $1 - \delta$ with fewer than $n(\delta, \epsilon)$ training examples.

Theorem 2: *any* algorithm that achieves loss less than ϵ with with probability greater than $1 - \delta$ requires more than $n(\delta, \epsilon)$ training examples.

Sample Mean, Sum of Gaussian RVs

- imagine $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$
- we're interested in the sample mean $\widehat{\mu} = \frac{1}{n}\sum_i X_i$

$$\widehat{\mu} \sim \mathcal{N}\left(\mu, \tfrac{\sigma^2}{n}\right)$$

- how good is our estimate μ?

$$\begin{aligned}
\mathbb{P}(|\widehat{\mu} - \mu| \geq t) &= 2\int_t^\infty \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} e^{-\frac{x^2}{2\sigma^2/n}}\, dx \\
&= 2\int_{t\sqrt{\frac{n}{\sigma^2}}}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \\
&= 2Q(t\sqrt{n/\sigma^2}) \\
&\leq 2e^{-t^2 n/2\sigma^2}
\end{aligned}$$

- upper bound for Gaussian tail

$$\begin{aligned}
Q(x) &:= \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \\
&\leq \int_x^\infty \frac{y}{x}\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \\
&= \frac{e^{-x^2/2}}{x\sqrt{2\pi}} \\
&\leq e^{-x^2/2} \text{ for } x \geq \frac{1}{\sqrt{2\pi}}
\end{aligned}$$

Sample Mean, Markov, Chebyshev

- we're interested in the sample mean $\widehat{\mu} = \frac{1}{n}\sum_i X_i$
- what if we don't know exact distribution of X_i?

Markov: if X is a non-negative RV, i.e, $X \geq 0$, then for any $t \geq 0$

$$\mathbb{P}(X \geq t) \leq \frac{E[X]}{t}$$

proof: $t\mathbb{I}\{X \geq t\} \leq X$

Chebyshev: Let $\widehat{\mu} = \frac{1}{n}\sum_i X_i$ and $\mu = E[X_i]$, then

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}$$

proof:
$$\begin{aligned}\mathbb{P}(|\widehat{\mu} - \mu| \geq \delta) &= \mathbb{P}((\widehat{\mu} - \mu)^2 \geq \delta^2) \\ &\leq \frac{E[(\widehat{\mu} - \mu)^2]}{\delta^2} \\ &= \frac{\mathrm{var}(\widehat{\mu})}{\delta^2} \\ &= \frac{\sigma^2}{n\delta^2}\end{aligned}$$

- the probability of $\widehat{\mu}$ deviating from $\mu \pm \delta$ is decreasing faster than $1/n$

Limit Results from Probability Theory

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean μ and variance σ.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum X_i - E[X_i]\right| \geq \delta\right) = 0 \qquad \text{for all } \delta > 0$$

Weak Law of Large Numbers
[convergence in probability]

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum X_i = E[X_i]\right) = 1$$

Strong Law of Large Numbers
[conv. with prob. 1 (wp1)]

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z) \quad \text{when} \quad Z_n := \frac{\sqrt{n}}{\sigma}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - E[X_i]\right]$$

$$\text{and } \Phi(z) := \int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx$$

Central Limit Theorem
[conv. in distribution]

Law of Large Numbers

Weak Law of Large Numbers. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean. Then $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to $E[X_i]$.

proof:

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum X_i - E[X_i]\right| \geq \delta\right) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\delta^2} \quad = 0 \qquad \text{for any } \delta > 0.$$

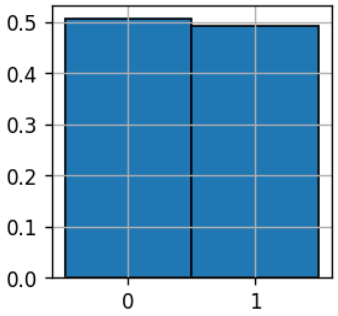
Strong Law of Large Numbers. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean. Then $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to $E[X_i]$.

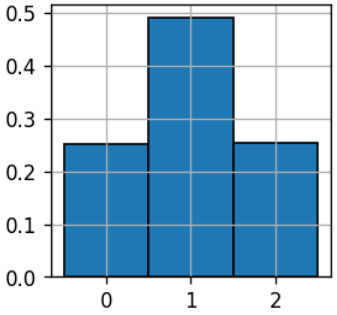
$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum X_i = E[X_i]\right) = 1$$

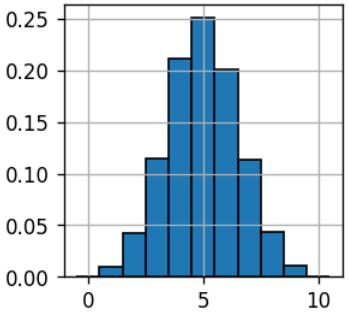
Central Limit Theorem

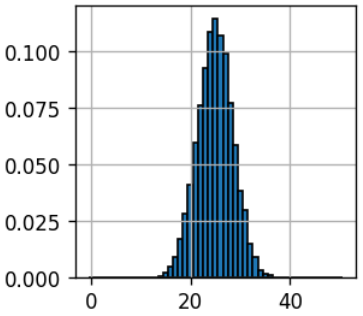
Central Limit Theorem. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean μ and finite variance σ^2. Then $\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

- example: $X_i \overset{i.i.d.}{\sim} \text{Bern}(0.5)$, $\sum_i X_i$ looks normal









- actual proof is a few lines:
  - express characteristic function of $\sqrt{n}(\widehat{\mu} - \mu)$ using a Taylor series
  - take limits and drop higher order terms

Example

- You get an internship for a political polling company. You find a way to sample population of voters at random, and observe that out of 1000 people polled, 550 will for party 1, and 450 plan to vote for party 2.

$$X_i \overset{i.i.d}{\sim} \text{Bernoulli}(\mu)$$

$$\widehat{\mu} = 0.55$$

Markov/Chebyshev:

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \geq \delta\right) \leq \frac{\sigma^2}{n\delta^2}$$

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \geq 0.05\right) \leq \frac{\sigma^2}{1000 \times 0.05^2} \leq 0.1$$

Example

• You get an internship for a political polling. You find a way to sample population of voters at random, and observe that out of 1000 people polled, 550 will for party 1, and 450 plan to vote for party 2.

$$X_i \overset{i.i.d}{\sim} \text{Bernoulli}(\mu)$$

$$\widehat{\mu} = 0.55$$

normal approximation:

$$\widehat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \geq \delta\right) = 2Q(\delta\sqrt{n/\sigma^2}) \approx 0.001565$$

exact:

$$\sum_i X_i \sim \text{Binomial}(k, 1000, \mu)$$

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \geq \delta\right) \approx 0.0015611$$

```python
import numpy as np
from scipy.stats import norm

two_sided_tail_norm = 2*(1-norm.cdf(0.05*np.sqrt(1000/.25)))
print('normal approximation: ', two_sided_tail_norm)
```

```
normal approximation:  0.0015654022580025018
```

```python
from scipy.stats import binom

exact_tail = 1-(binom.cdf(550, 1000, .5)-binom.cdf(450,1000,.5))
print('exact tail:', exact_tail)
```

```
exact tail: 0.0015611388396998827
```

Hoeffding's Inequality

Hoeffding's Inequality. Let $X_1, X_2, \ldots, X_n$ be i.i.d. bounded random variables with $X_i \in [a, b]$ and define $\widehat{\mu} = \frac{1}{n}\sum X_i$, $\mu = E[X_i]$. Then

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq 2e^{-\frac{2t^2 n}{(b-a)^2}}$$

- example

$$X_i \overset{i.i.d}{\sim} \text{Bernoulli}(\mu)$$

$$\widehat{\mu} = 0.55$$

Hoeffding

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq 2e^{-\frac{2nt^2}{(b-a)^2}}$$

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq 0.05) \leq 2e^{-2\times 1000\times 0.05^2} = 2e^{-5} \quad \approx 0.0134$$

Hoeffding's Inequality

Hoeffding's Inequality. Let $X_1, X_2, \ldots, X_n$ be i.i.d. bounded random variables with $X_i \in [a, b]$ and define $\widehat{\mu} = \frac{1}{n}\sum X_i$, $\mu = E[X_i]$. Then

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq 2e^{-\frac{2t^2 n}{(b-a)^2}}$$

- imagine $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq 2e^{-t^2 n/\sigma^2}$$

- Hoeffding is essentially as good as knowing the distributions are normal!
- proof of Hoeffding uses Chernoff and carefully selecting s (see note)

$$\mathbb{P}(X \geq t) \leq e^{-st} E[e^{sX}]$$

$$\mathbb{P}\left(\widehat{\mu} - \mu \geq t\right) \leq e^{-st} E[e^{s(\widehat{\mu}-\mu)}]$$

$$\vdots$$

Chernoff Bound

Chernoff Bound (Cramer, 1938). For a random variable X and any $s \geq 0$, $t \in \mathbb{R}$, $\mathbb{P}(X \geq t) \leq e^{-st} E[e^{sX}]$.

proof:

$$\mathbb{I}\{X \geq t\} \leq e^{s(X-t)} \qquad \text{for } s \geq 0$$

Gaussian Tail Bound Using Chernoff

Classic Gaussian Tail bound. Let $X \sim \mathcal{N}(0,1)$. Then for $t \geq 0$,

$$\mathbb{P}(X \geq t) \leq e^{-t^2/2}$$

proof:

$$\begin{aligned}\mathbb{P}(X \geq t) &\leq e^{-st} E[e^{sX}] \\ &= e^{-st} e^{s^2/2} \\ &\leq e^{-t^2/2}\end{aligned}$$

by choosing $s = t$

- moment generating function $E[e^{sX}]$ for $\mathcal{N}(0,1)$:

$$E[e^{sX}] = e^{s^2/2}$$