Introduction to Machine Learning
Lecture 2

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Basic Probability Notions

Probabilistic Model

- Sample space: Ω, set of all outcomes or elementary events possible in a trial, e.g., casting a die or tossing a coin.

- Event: subset $A \subseteq \Omega$ of sample space. The set of all events must be closed under complementation and countable union and intersection.

- Probability distribution: mapping $\Pr$ from the set of all events to $[0, 1]$ such that $\Pr[\Omega] = 1$, and for all mutually exclusive events,

$$\Pr[A_1 \cup \ldots \cup A_n] = \sum_{i=1}^{n} \Pr[A_i].$$

Random Variables

- Definition: a random variable is a function $X\colon \Omega \rightarrow \mathbb{R}$ such that for any interval I, the subset of the sample space $\{A\colon X(A) \in I\}$ is an event. Such a function is said to be measurable.
- Example: the sum of the values obtained when casting a die.
- Probability mass function of random variable X: function $f\colon x \mapsto f(x) = \Pr[X = x]$.
- Joint probability mass function of X and Y:

$$f\colon (x, y) \mapsto f(x, y) = \Pr[X = x \wedge Y = y].$$

Conditional Probability and Independence

- Conditional probability of event A given B:

$$\Pr[A \mid B] = \frac{\Pr[A \wedge B]}{\Pr[B]},$$

when $\Pr[B] \neq 0.$

- Independence: two events A and B are independent when

$$\Pr[A \wedge B] = \Pr[A] \Pr[B].$$

Equivalently, $\Pr[A \mid B] = \Pr[A]$, when $\Pr[B] \neq 0.$

Some Probability Formulae

- Sum rule:

$$\Pr[A \vee B] = \Pr[A] + \Pr[B] - \Pr[A \wedge B].$$

- Union bound:

$$\Pr[\bigvee_{i=1}^{n} A_i] \leq \sum_{i=1}^{n} \Pr[A_i].$$

- Bayes formula:

$$\Pr[X \mid Y] = \frac{\Pr[Y \mid X] \Pr[X]}{\Pr[Y]} \quad (\Pr[Y] \neq 0).$$

Some Probability Formulae

- Chain rule:

$$\Pr[\bigwedge_{i=1}^{n} X_i] = \Pr[X_1] \Pr[X_2 \mid X_1] \Pr[X_3 \mid X_1 \wedge X_2] \\ \ldots \Pr[X_n \mid \bigwedge_{i=1}^{n-1} X_i].$$

- Theorem of total probability: assume that

$$\Omega = A_1 \cup A_2 \cup \ldots \cup A_n, \ \text{with } A_i \cap A_j = \emptyset \text{ for } i \neq j;$$

then for any event B,

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B \mid A_i] \Pr[A_i].$$

Expectation

- Definition: the *expectation* (or *mean*) of a random variable X is

$$\mathrm{E}[X] = \sum_x x \Pr[X = x].$$

- Properties:
  - linearity, $\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$.
  - if X and Y are independent,

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

Expectation

- **Theorem** (Markov's inequality): let X be a non-negative random variable with $\mathrm{E}[X] < \infty$, then for all $t > 0$,

$$\Pr[X \geq t\mathrm{E}[X]] \leq \frac{1}{t}.$$

- **Proof**: $$\begin{aligned}\Pr[X \geq t\,\mathrm{E}[X]] &= \sum_{x \geq t\mathrm{E}[X]} \Pr[X = x] \\ &\leq \sum_{x \geq t\,\mathrm{E}[X]} \Pr[X = x] \frac{x}{t\,\mathrm{E}[X]} \\ &\leq \sum_{x} \Pr[X = x] \frac{x}{t\,\mathrm{E}[X]} \\ &= \mathrm{E}\left[\frac{X}{t\,\mathrm{E}[X]}\right] = \frac{1}{t}.\end{aligned}$$

Variance

- Definition: the *variance* of a random variable X is

$$\mathrm{Var}[X] = \sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2].$$

σ_X is called the *standard deviation* of the random variable X.

- Properties:
  - $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X].$
  - if X and Y are independent,

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Variance

- Theorem (Chebyshev's inequality): let X be a random variable with $\mathrm{Var}[X] < \infty$, then for all $t > 0$,

$$\Pr[|X - \mathrm{E}[X]| \geq t\sigma_X] \leq \frac{1}{t^2}.$$

- Proof: Observe that

$$\Pr[|X - \mathrm{E}[X]| \geq t\sigma_X] = \Pr[(X - \mathrm{E}[X])^2 \geq t^2\sigma_X^2].$$

The result follows Markov's inequality.

Application

- Experiment: roll a pair of fair dice *n* times. Can we give a good estimate of the sum of the values after *n* rolls?

- Mean: *7n*, variance: *35/6 n*; thus by Chebyshev's inequality, the final sum will lie between

$$7n - 10\sqrt{\frac{35}{6}n} \text{ and } 7n + 10\sqrt{\frac{35}{6}n}$$

in at least *99%* of all experiments. The odds are better than *99* to *1* that the sum be roughly between *6.976M* and *7.024M* after *1M* rolls.

Weak Law of Large Numbers

- **Theorem**: let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with the same mean μ and variance $\sigma^2 < \infty$ and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then for any $\epsilon > 0$, $\lim_{n\to\infty} \Pr[|\overline{X}_n - \mu| \geq \epsilon] = 0.$
- **Proof**: Since the variables are independent,

$$\mathrm{Var}[\overline{X}_n] = \sum_{i=1}^{n} \mathrm{Var}\left[\frac{X_i}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

- Thus, by Chebyshev's inequality,

$$\Pr[|\overline{X}_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Hoeffding's Theorem

- **Theorem**: Let $X_1, \ldots, X_m$ be independent random variables $X_i \in [a_i, b_i]$. Then for $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^m X_i$:

$$\Pr[S_m - \mathrm{E}[S_m] \geq \epsilon] \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2}$$
$$\Pr[S_m - \mathrm{E}[S_m] \leq -\epsilon] \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2}.$$

- **Proof**: The proof is based on Chernoff's bounding technique: for any random variable X and $t > 0$, apply Markov's inequality and select t to minimize

$$\Pr[X \geq \epsilon] = \Pr[e^{tX} \geq e^{t\epsilon}] \leq \frac{\mathrm{E}[e^{tX}]}{e^{t\epsilon}}.$$

- Using this scheme and the independence of the random variables gives

$$
\begin{aligned}
\Pr[S_m - \mathrm{E}[S_m] \geq \epsilon] & \\
& \leq e^{-t\epsilon}\, \mathrm{E}[e^{t(S_m - \mathrm{E}[S_m])}] \\
& = e^{-t\epsilon} \Pi_{i=1}^m \, \mathrm{E}[e^{t(X_i - \mathrm{E}[X_i])}] \\
(\text{lemma applied to } X_i - \mathrm{E}[X_i]) \; & \leq e^{-t\epsilon} \Pi_{i=1}^m e^{t^2 (b_i - a_i)^2/8} \\
& = e^{-t\epsilon} e^{t^2 \sum_{i=1}^m (b_i - a_i)^2/8} \\
& \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2},
\end{aligned}
$$

choosing $t = 4\epsilon / \sum_{i=1}^m (b_i - a_i)^2$.

- The second inequality is proved in a similar way.

Hoeffding's Lemma

- **Lemma**: Let X be a random variable with $\mathrm{E}[X] = 0$ and $a \leq X \leq b$ with $b \neq a$. Then for $t > 0$,

$$\mathrm{E}[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}.$$

- **Proof**: by convexity of $x \mapsto e^{tx}$, for all $a \leq x \leq b$,

$$e^{tx} \leq \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}.$$

Thus,

$$\mathrm{E}[e^{tX}] \leq \mathrm{E}[\tfrac{b-X}{b-a} e^{ta} + \tfrac{X-a}{b-a} e^{tb}] = \tfrac{b}{b-a} e^{ta} + \tfrac{-a}{b-a} e^{tb} = e^{\phi(t)},$$

with,

$$\phi(t) = \log(\tfrac{b}{b-a} e^{ta} + \tfrac{-a}{b-a} e^{tb}) = ta + \log(\tfrac{b}{b-a} + \tfrac{-a}{b-a} e^{t(b-a)}).$$

- Taking the derivative gives:

$$\phi'(t) = a - \frac{ae^{t(b-a)}}{\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}} = a - \frac{a}{\frac{b}{b-a}e^{-t(b-a)} - \frac{a}{b-a}}.$$

- Note that: $\phi(0) = 0$ and $\phi'(0) = 0$. Furthermore,

$$\begin{aligned}
\Phi''(t) &= \frac{-abe^{-t(b-a)}}{[\frac{b}{b-a}e^{-t(b-a)} - \frac{a}{b-a}]^2} \\
&= \frac{\alpha(1-\alpha)e^{-t(b-a)}(b-a)^2}{[(1-\alpha)e^{-t(b-a)} + \alpha]^2} \\
&= \frac{\alpha}{[(1-\alpha)e^{-t(b-a)} + \alpha]} \frac{(1-\alpha)e^{-t(b-a)}}{[(1-\alpha)e^{-t(b-a)} + \alpha]}(b-a)^2 \\
&= u(1-u)(b-a)^2 \leq \frac{(b-a)^2}{4},
\end{aligned}$$

with $\alpha = \frac{-a}{b-a}$. There exists $0 \leq \theta \leq t$ such that:

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\theta) \leq t^2 \frac{(b-a)^2}{8}.$$

Example: Tossing a Coin

- **Problem**: estimate bias p of a coin.

$$\text{H, T, T, H, T, H, H, T, H, H, H, T, T, } \ldots \text{, H.}$$

- Let $h = 1_{\text{H}}$. Then $p = R(h)$ and $\widehat{p} = \widehat{R}(h)$ is the percentage of tails in the sample. Thus, with probability at least $1 - \delta$,

$$|p - \widehat{p}| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Thus, choosing $\delta = .02$ and $m = 1000$ implies that with probability at least 98%,

$$|p - \widehat{p}| \leq \sqrt{\log(10)/1000} \approx .048.$$

McDiarmid's Inequality

(McDiarmid, 1989)

- **Theorem**: let $X_1, \ldots, X_m$ be independent random variables taking values in U and $f : U^m \to \mathbb{R}$ a function verifying for all $i \in [1, m]$,

$$\sup_{x_1, \ldots, x_m, x_i'} |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c.$$

Then, for all $\epsilon > 0$,

$$\Pr\Big[\big|f(X_1, \ldots, X_m) - \mathrm{E}[f(X_1, \ldots, X_m)]\big| > \epsilon\Big] \leq 2 \exp\left(-\frac{2\epsilon^2}{mc^2}\right).$$