Aprendizado Estatístico

Semana 2 - Concentração de medida

Bruno Marcondes e Resende e Felipe Luis Giacomini

Na aula anterior, estudamos as desigualdades de Markov, Chebyshev e Chernoff. Nesta aula, nosso objetivo é demonstrar o lema de Hoeffding, enunciado a seguir.

Lema 1.1. *(Lema de Hoeffding) Seja X uma variável aleatória tal que $X \in [a, b]$. Então, para todo $t \in \mathbb{R}$, temos que*

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

Demonstração. O método que usaremos para provar o lema irá resultar em um limitante superior com uma constante um pouco pior no expoente, porém, ele tem a vantagem de ser mais legal. Para isso, considere σ uma variável aleatória tal que $\mathbb{P}(\sigma = 1) = \mathbb{P}(\sigma = -1) = 1/2$, chamada de **variável de Rademacher**. Usando a série de Taylor da função exponencial, temos que

$$\begin{aligned}
\mathbb{E}\left[e^{t\sigma}\right] &= \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(t\sigma)^n}{n!}\right] \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} t^n \mathbb{E}\left[\sigma^n\right] \\
&= \sum_{n=0}^{\infty} \frac{t^{2n}}{(2n)!} \\
&\leq \sum_{n=0}^{\infty} \frac{t^{2n}}{n!\, 2^n} \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{t^2}{2}\right)^n \\
&= e^{t^2/2}.
\end{aligned}$$

No desenvolvimento acima usamos que, para todo n natural, temos

$$\mathbb{E}\left[\sigma^n\right] = \begin{cases} 0, & \text{se n é ímpar} \\ 1, & \text{se n é par.} \end{cases}$$

Além disso,

$$(2n)! = \underbrace{(2n)(2n-1)\dots(n+1)}_{\geq 2^n} n \dots 2 \cdot 1 \geq 2^n n!\,.$$

Assim, concluímos que o lema é válido para a variável σ. Para estender o resultado para variáveis limitadas em geral, aplicaremos um método bastante conhecido em probabilidade, chamado de simetrização. Para isso, usaremos sem provar a desigualdade de Jensen, segundo a qual para $f : \mathbb{R} \to \mathbb{R}$ uma função convexa e X uma variável aleatória, temos

$$\mathbb{E}\left[f(X)\right] \geq f(\mathbb{E}\left[X\right]).$$

Com isso, seja $X \in [a,b]$ uma variável aleatória e X' independente de X e com a mesma distribuição. Então

$$\begin{aligned}
\mathbb{E}_X\left[\exp(t(X-\mathbb{E}_X\left[X\right]))\right] &= \mathbb{E}_X\left[\exp(t(X-E_{X'}\left[X'\right]))\right] \\
&= \mathbb{E}\left[\exp(\mathbb{E}_{X'}\left[t(X-X')\right])\right] \\
&\leq \mathbb{E}_X\left[\mathbb{E}_{X'}\left[\exp(t(X-X'))\right]\right] \\
&= \mathbb{E}_{X,X'}\left[\exp(t(X-X'))\right].
\end{aligned}$$

Note que $X-X'$ tem a mesma distribuição de $X'-X$ e, portanto, $\sigma(X-X')$ tem a mesma distribuição de $X-X'$. Assim, usando as propriedades de esperança condicional, obtemos

$$\begin{aligned}
\mathbb{E}_{XX'}\left[\exp(t(X-X'))\right] &= \mathbb{E}_{XX',\sigma}\left[\exp(t\sigma(X-X'))\right] \\
&= \mathbb{E}_{XX'}\left[\mathbb{E}_\sigma\left[\exp(t\sigma(X-X'))|X,X'\right]\right] \\
&\leq \mathbb{E}_{XX'}\left[\exp\left(\frac{t^2(X-X')^2}{2}\right)\right] \\
&\leq e^{t^2(b-a)^2/2},
\end{aligned}$$

em que na primeira desigualdade aplicamos o resultado já demonstrado para σ e na segunda o fato de $X-X'$ ser no máximo $b-a$. Este é o resultado com um fator de 2 em vez de 8 no expoente, a demonstração está completa. $\square$

Teorema 1.1. *(Desigualdade de Hoeffding) Se $X \in [a,b]$, então:*

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}(X_i-\mathbb{E}[X]) \geq \varepsilon\right) \leq \exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right)$$

Demonstração. Usaremos o limitante de Chernoff para obter o resultado apresentado.

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}(X_i-\mathbb{E}[X]) \geq \varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^{m}(X_i-\mathbb{E}[X]) \geq m\varepsilon\right) \leq \frac{\mathbb{E}\left(e^{t\sum_{i=1}^{m}(X_i-\mathbb{E}[X])}\right)}{e^{t\varepsilon m}}$$

$$= \frac{1}{e^{t\varepsilon m}}\prod_{i=1}^{m}\mathbb{E}\left(e^{t(X_i-\mathbb{E}[X])}\right) \leq \frac{1}{e^{t\varepsilon m}}\prod_{i=1}^{m}\left(e^{\frac{t^2(b-a)^2}{2}}\right) = e^{\frac{mt^2(b-a)^2}{2}-t\varepsilon m}, \quad \forall\, t>0.$$

Agora precismos minimizar em t. Isto equivale a minimizar a seguinte função:

$$f(t) = \frac{mt^2(b-a)^2}{2} - t\,\varepsilon m.$$

Derivando e igualando a 0, obtemos

$$\begin{aligned}
f'(t) &= m(b-a)^2 t - \varepsilon m = 0 \\
\Rightarrow \quad t &= \frac{\varepsilon}{m(b-a)^2}.
\end{aligned}$$

Vamos substituir de volta na última desigualdade.

$$\leq \exp\left(\frac{m\,\varepsilon^2\,(b-a)^2}{2(b-a)^4} - \frac{\varepsilon}{(b-a)^2}\,\varepsilon\, m\right) = \exp\left(\frac{m\,\varepsilon^2}{2\,(b-a)^2} - \frac{m\varepsilon^2}{(b-a)^2}\right) = \exp\left(-\frac{\varepsilon^2\, m}{2(b-a)^2}\right).$$

Portanto, mostramos que

$$\mathbb{P}\left(\frac{1}{m}\sum X_i - \mathbb{E}[X_i] \geq \varepsilon\right) \leq e^{\frac{-\varepsilon^2 m}{2(b-a)^2}}.$$

$\square$

No entanto, o resultado que obtivemos não é ótimo. O resultado ótimo seria: $e^{-\frac{2\varepsilon^2 m}{(b-a)^2}}$. No que segue, iremos usar o resultado ótimo.

Note que, tomando $Z_i = -X_i$, temos $Z_i \in [-b, -a]$. Então, podemos usar o resultado acima para essa variável aleatória.

$$\mathbb{P}\left(\frac{1}{m}\sum Z_i - \mathbb{E}[Z_i] \geq \varepsilon\right) \leq e^{\frac{-2\varepsilon^2 m}{(b-a)^2}}.$$

Note que

$$\mathbb{P}\left(\frac{1}{m}\sum Z_i - \mathbb{E}[Z_i] \geq \varepsilon\right) = \mathbb{P}\left(-\frac{1}{m}\sum X_i + \mathbb{E}[X_i] \geq \varepsilon\right) = \mathbb{P}\left(\frac{1}{m}\sum X_i - \mathbb{E}[X_i] \leq -\varepsilon\right).$$

Logo,

$$\mathbb{P}\left(\frac{1}{m}\sum X_i - \mathbb{E}[X_i] \leq -\varepsilon\right) \leq e^{\frac{-2\varepsilon^2 m}{(b-a)^2}}.$$

Então, isso implica que

$$\mathbb{P}\left(\left|\frac{1}{m}\sum X_i - \mathbb{E}[X_i]\right| \geq \varepsilon\right) \leq 2e^{\frac{-2\varepsilon^2 m}{(b-a)^2}}.$$

Em particular, a v.a. no nosso caso é $X_i = \mathbb{I}(h(x_i) \neq y_i) \in [0, 1]$.

Com esse resultado consguimos provar só para um $h \in \mathcal{H}$. Para provar para todos, usaremos o Union Bound.

$$\begin{gathered}
\mathbb{P}\left(\exists h, |L_S(h) - L_D(h)| \geq \varepsilon\right) \\
= \mathbb{P}\left(|L_s(h_1) - L_D(h_1)| \geq \varepsilon \,\cup\, |L_s(h_2) - L_D(h_2)| \geq \varepsilon \,\cup\, \ldots \,\cup\, |L_s(h_{|\mathcal{H}|}) - L_D(h_{|\mathcal{H}|})| \geq \varepsilon\right) \\
\leq \sum_{h \in \mathcal{H}} \mathbb{P}(|L_s(h) - L_D(h)| \geq \epsilon) \leq |\mathcal{H}| 2e^{\frac{-2\varepsilon^2 m}{(b-a)^2}}.
\end{gathered}$$

Para achar o tamanho amostral, basta isolar o m em função dos demais valores.

$$\begin{aligned}
\delta &\sim |\mathcal{H}| 2e^{-2\varepsilon^2 m} \\
\frac{\delta}{2} &\sim |\mathcal{H}| e^{-2\varepsilon^2 m} \\
\log\left(\frac{\delta}{2}\right) &\sim \log(|\mathcal{H}|) - 2\varepsilon^2 m \\
\Rightarrow \quad m &\sim \frac{1}{2\varepsilon^2}\left(\log(|\mathcal{H}|) + \log\left(\frac{2}{\delta}\right)\right).
\end{aligned}$$

Então, com probabilidade $1-\delta$, seu modelo não sofre de overfitting (a menos de ε) se o seu tamanho amostral for maior ou igual a $\frac{1}{2\varepsilon^2}\left(\log(|\mathcal{H}|)+\log\left(\frac{2}{\delta}\right)\right), \forall\, \varepsilon, \delta > 0$

Definição 1.1. *(Growth Function): A função de crescimento* $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ *é dada por*

$$\Pi_{\mathcal{H}}(m) = \max_{x_1,\ldots,x_m \in \mathcal{X}} \left|\{(h(x_1),\ldots,h(x_m)) : h \in \mathcal{H}\}\right|.$$

$\Pi_{\mathcal{H}}$ *é o máximo número de formas distintas que* m *pontos podem ser classificados por* $\mathcal{H}$*.*