Aprendizado Estatístico

Desigualdades de Concentração: McDiarmid

Gabriela Ishida e Vagner Santos

Desigualdade de McDiarmid

A **Desigualdade de McDiarmid**, introduzida pelo matemático britânico Colin McDiarmid em 1989, é uma ferramenta poderosa na teoria das probabilidades e estatística, especialmente no estudo de fenômenos estocásticos e concentrações de medida. A motivação principal para sua formulação está ligada à análise de estabilidade de funções que dependem de variáveis aleatórias independentes.

1.0.1 História e Motivação

A desigualdade de McDiarmid surgiu no contexto da teoria das probabilidades aplicada a problemas de aprendizado de máquina, otimização combinatória, e outros campos onde é essencial entender como pequenas perturbações em variáveis aleatórias influenciam a saída de uma função. Uma questão-chave que motivou McDiarmid foi a necessidade de generalizar outras desigualdades de concentração (como as desigualdades de Hoeffding) para funções mais gerais que dependem de múltiplas variáveis.

A desigualdade foi desenvolvida para medir a robustez ou estabilidade de funções que não são sensíveis a alterações pequenas em seus argumentos — uma propriedade chamada *bounded differences* (diferenças limitadas). Essa propriedade estabelece que mudar o valor de uma única variável de entrada afeta a saída da função apenas dentro de um limite conhecido.

1.0.2 Uso e Aplicações

A desigualdade de McDiarmid tem um papel central em diversas áreas da ciência e tecnologia:

1. **Teoria do Aprendizado de Máquina**: Avaliação do risco empírico em problemas de aprendizado estatístico. Ela fornece limites de concentração para a diferença entre o risco empírico e o risco esperado, garantindo a generalização de modelos treinados.

2. **Otimização Combinatória**: Análise de problemas complexos como o corte mínimo, fluxo máximo e problemas de roteamento, onde as soluções dependem de várias variáveis independentes.

3. **Teoria da Informação**: Avaliação da robustez de sistemas que processam dados sujeitos a ruído ou incertezas.

4. **Redes e Sistemas Complexos**: Estudo da estabilidade de redes quando submetidas a mudanças locais em seus parâmetros.

5. **Análise Estatística**: Controle da probabilidade de desvios de uma variável aleatória composta de sua expectativa.

1.0.3 Ideia Central

A desigualdade afirma que, se uma função satisfaz a propriedade de diferenças limitadas, então a probabilidade de um desvio significativo em relação à sua média (ou valor esperado) decai exponencialmente à medida que o desvio cresce. Isso oferece garantias matemáticas para muitos sistemas dependentes de variáveis aleatórias independentes.

Definição 1.1 (Desigualdade de McDiarmid)**.** *: Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes, onde X_i pertence ao espaço X_i. Seja $f : X_1 \times \cdots \times X_n \to \mathbb{R}$ uma função com a propriedade de diferenças limitadas $(c_1, \ldots, c_n)$: para todo $i = 1, \ldots, n$ e para todos $(x_1, x_2, \ldots, x_i, x_{i+1} \ldots, x_n), (x_1, x_2, \ldots, x_{i-1}, y, x_{i+1} \ldots, x_n) \in X_1 \times \cdots \times X_n$, que diferem apenas na i-ésima coordenada ($x_j = y$ para todo $j \neq i$), temos que:*

$$\sup_{y \in \mathbb{R}} |f(x_1, x_2, \ldots, x_i, x_{i+1} \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, y, x_{i+1} \ldots, x_n)| \leq c_i.$$

Então, para qualquer $\xi > 0$,

$$\mathbb{P}(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq \xi) \leq \exp\left(-\frac{2\xi^2}{\sum_{i=1}^n c_i^2}\right).$$

Demonstração:

Escrevemos $\mathbb{E}_i[\cdot]$ para denotar a expectativa condicionada a $X_{1:i} := (X_1, \ldots, X_i)$ e, $\mathbf{X} := (X_1, X_2, \ldots, X_n)$. Assim, definimos:

$$\mathbb{E}_i[f(\mathbf{X})] := \mathbb{E}[f(X_1, \ldots, X_n) \mid X_{1:i}]$$

Alguns resultados importantes:

- $\mathbb{E}_0[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{X})]$,
- $\mathbb{E}_n[f(\mathbf{X})] = f(\mathbf{X})$.

E definimos as variáveis aleatórias:

$$V_i := \mathbb{E}_i[f(\mathbf{X})] - \mathbb{E}_{i-1}[f(\mathbf{X})],$$

$$A_i := \inf_{x_i \in X_i} (\mathbb{E}_i[f(X_{1:i-1}, x_i)] - \mathbb{E}_{i-1}[f(\mathbf{X})]),$$

$$B_i := \sup_{x_i \in X_i} (\mathbb{E}_i[f(X_{1:i-1}, x_i)] - \mathbb{E}_{i-1}[f(\mathbf{X})]).$$

Essas variáveis satisfazem:

$$V_i \in [A_i, B_i],$$

$$\mathbb{E}_{i-1}[V_i] = 0,$$

$$\begin{aligned}\sum_{i=1}^n V_i &= (\mathbb{E}_1[f(\mathbf{X})] - \mathbb{E}_0[f(\mathbf{X})]) + (\mathbb{E}_2[f(\mathbf{X})] - \mathbb{E}_1[f(\mathbf{X})]) + \ldots + (\mathbb{E}_n[f(\mathbf{X})] - \mathbb{E}_{n-1}[f(\mathbf{X})]) \\ &= -\mathbb{E}_0[f(\mathbf{X})] + \mathbb{E}_n[f(\mathbf{X})] \\ &= f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})].\end{aligned}$$

Reivindicamos que $[A_i, B_i]$ é sempre um intervalo de comprimento no máximo c_i. A demonstração será feita mais tarde.

Agora, para limitar a probabilidade de que a soma dos V_i seja pelo menos ξ, usamos a desigualdade de Chernoff. Definimos $S_i := V_1 + \cdots + V_i$. Assim:

$$\begin{aligned}
\mathbb{P}(f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n) \geq \xi) &= \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq \xi\right) \\
&= \mathbb{P}(S_n \geq \xi) \\
&\leq e^{-\lambda\xi}\mathbb{E}[e^{\lambda S_n}] = e^{-\lambda\xi}\mathbb{E}[e^{\lambda(V_n+S_{n-1})}]. \quad \text{(Chernoff)}
\end{aligned}$$

Usando a propriedade de condicionamento:

$$\begin{aligned}
\mathbb{E}[e^{\lambda S_n}] &= \mathbb{E}[e^{\lambda(V_n+S_{n-1})}] \\
&= \mathbb{E}[\mathbb{E}[e^{\lambda V_n} e^{\lambda S_{n-1}} \mid X_{1:(n-1)}]] \\
&= \mathbb{E}[\mathbb{E}[e^{\lambda V_n} \mid X_{1:(n-1)}] e^{\lambda S_{n-1}}] \\
&= \mathbb{E}[\mathbb{E}_{n-1}[e^{\lambda V_n}] e^{\lambda S_{n-1}}]
\end{aligned}$$

Aplicando a desigualdade de Hoeffding:

$$\mathbb{E}_{n-1}[e^{\lambda V_n}] \leq e^{\frac{\lambda^2 c_n^2}{8}},$$

Com esses desenvolvimentos, temos:

$$\mathbb{E}[e^{\lambda S_n}] \leq e^{\frac{\lambda^2 c_n^2}{8}} \mathbb{E}[e^{\lambda S_{n-1}}].$$

Note que

$$\mathbb{E}[e^{\lambda S_{n-1}}] \leq e^{\frac{\lambda^2 c_{n-1}^2}{8}} \mathbb{E}[e^{\lambda S_{n-2}}].$$

Repetindo o mesmo desenvolvimento para $S_{n-2}, S_{n-3}, \ldots S_2, S_1$, obtemos:

$$\mathbb{E}[e^{\lambda S_n}] \leq \prod_{i=1}^{n} e^{\frac{\lambda^2 c_i^2}{8}} = e^{\frac{\lambda^2}{8}\sum_{i=1}^{n} c_i^2}. \tag{1.1}$$

Com (1.1), temos que

$$\begin{aligned}
\mathbb{P}(f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n) \geq \xi) &= \mathbb{P}(S_n \geq \xi) \\
&\leq e^{-\lambda\xi} e^{\frac{\lambda^2}{8}\sum_{i=1}^{n} c_i^2}.
\end{aligned}$$

Usando o λ ótimo, dado pelo $\arg\min_{\lambda>0}\{e^{-\lambda\xi} e^{\frac{\lambda^2}{8}\sum_{i=1}^{n} c_i^2}\} = \dfrac{4\xi}{\sum_{i=1}^{n} c_i^2}$

$$\mathbb{P}(f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n) \geq \xi) \leq \exp\left(-\frac{2\xi^2}{\sum_{i=1}^{n} c_i^2}\right). \quad \square$$

Comprimento do intervalo $[A_i, B_i]$

Resta provar que $[A_i, B_i]$ é um intervalo de comprimento no máximo c_i. Por independência de $X_1, \ldots, X_n$, temos:

$$\mathbb{E}_i[f(X_{1:i-1}, x_i)] = \mathbb{E}[f(X_{1:i-1}, x_i, X_{i+1:n}) \mid X_{1:i-1}].$$

Logo:

$$B_i - A_i = \sup_{b \in X_i} \mathbb{E}_i[f(X_{1:i-1}, b)] - \inf_{a \in X_i} \mathbb{E}_i[f(X_{1:i-1}, a)].$$

Usando a propriedade de diferenças limitadas de f:

$$B_i - A_i \leq c_i.$$

1.1 Exemplos

Exemplo 1.2. *Exemplo: Bootstrap*

$$F(X_1, \ldots, X_n) = \sup_{t \in \mathbb{R}} \left| \hat{P}_n(X \leq t) - P(X \leq t) \right|$$

$$\hat{P}_n(X \leq t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}$$

Se $F(X_1, \ldots, X_i', \ldots, X_n) = F(X_1, \ldots, X_n)$*, temos:*

$$F(X_1, \ldots, X_i', \ldots, X_n) - F(X_1, \ldots, X_n) \leq \frac{1}{n}$$

Utilizando a desigualdade de McDiarmid, obtemos:

$$P\left(\sup_t \left| \hat{P}_n(X \leq t) - P(X \leq t) \right| > \varepsilon \right) \leq e^{-\frac{2n\varepsilon^2}{1}}$$

A ideia é usar a propriedade de dimensão VC.