MATHEMATICS OF DATA SCIENCE

November 14, 2023

Today: Statistics for data science
5 lectures, 3 tutorials + homework

STATISTICS AND CONCENTRATION INEQUALITIES

Motivation: Probabilistic reasoning is common in data science

→ Random models of data: even when the data is deterministic, it is often helpful to think of the data as originating from some distribution

→ Randomized algorithms: the most effective to perform data science tasks, based on statistical principles

Our focus: Concentration inequalities

① Basics of probability (scalar variables)

↳ Probability space: $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ $\quad$ $\mathbb{R}$: real numbers $(-\infty, \infty)$ "universe"

$\mathcal{B}(\mathbb{R})$: Borel σ-algebra (set of events)

$\mathbb{P}: \mathcal{B}(\mathbb{R}) \to [0,1]$ probability measure

$\forall B \in \mathcal{B}(\mathbb{R}), \ \mathbb{P}(B) \in [0,1]$ $\quad$ $\mathbb{P}(\mathbb{R}) = 1$, $\mathbb{P}(\emptyset) = 0$

↳ A random variable Y can be seen as a function from $\mathcal{B}(\mathbb{R})$ to $[0,1]$ defined for any $B \in \mathcal{B}(\mathbb{R})$ $(B \subseteq \mathbb{R})$ by $\mathbb{P}(Y \in B)$

. Discrete random variables: $\{y_i\}_{i \in S}$, S is finite or countable

$\{p_i\}_{i\in S}$ such that $p_i \geq 0$ $\sum_{i\in S} p_i = 1$

$\mathcal{Y}$ set of possible values for y

$\forall A \subseteq \mathcal{Y} = \{y_i\}_{i\in S}$, $\mathbb{P}(Y \in A) = \sum_{y_i \in A} p_i = \sum_{y_i \in A} \mathbb{P}(Y = y_i)$

$\left(\forall B \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}(Y \in B) = \sum_{y_i \in B} p_i \right)$

Y : random variable $\qquad \{y_i\}_{i\in S}$: fixed, possible values for Y

Example: Roll a dice

Y : value of the roll $\in \mathbb{R}$

$\mathcal{Y} = \{1, 2, 3, 4, 5, 6\}$ $\qquad p_1 = \dots = p_6 = \frac{1}{6}$

$\mathbb{P}(Y \in \{1, 2, 3\}) = p_1 + p_2 + p_3 = \frac{1}{2}$

$\mathbb{P}(Y \in [0, 3.5]) = \frac{1}{2}$

- Continuous random variables

Y random variable $\qquad \mathcal{Y} \subseteq \mathbb{R}$ with possibly infinite cardinality (e.g. $\mathcal{Y} = [0,1]$)

$p : \mathcal{Y} \to \mathbb{R}$

probability density function

$p(y) \geq 0 \ \forall y \in \mathcal{Y}$ and $\int_{\mathcal{Y}} p(y)\,dy = 1$

$\forall A \subseteq \mathcal{Y}, \quad \mathbb{P}(y \in A) = \int_A p(y)\,dy$

Ex) | Gaussian random variable (μ, σ^2) $\mu \in \mathbb{R}$, $\sigma > 0$
Normal

$$\mathcal{Y} = \mathbb{R} \qquad p(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

Gaussian / Normal density

Standard normal variable : $\mu = 0$, $\sigma = 1$

Notation : $Y \sim \mathcal{N}(\mu, \sigma^2)$: Y is a random variable with
· Gaussian density defined by μ and σ^2

$Y \sim p$: Y is a random variable with density p

$Y \sim \{p_i\}_{i \in S}$: ______ discrete
probabilities $\{p_i\}$

Expected value / Mean

If Y is a random variable, then

$$\mathbb{E}[Y] = \begin{cases} \sum_{y_i \in \mathcal{Y}} y_i\, p_i & \text{if } Y \text{ discrete} \\ \int_{\mathcal{Y}} y\, p(y)\, dy & \text{if } Y \text{ continuous} \end{cases}$$

Ex) $Y \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[Y] = \mu$

Property: $\forall (\alpha,\beta) \in \mathbb{R}^2, \; \mathbb{E}[\alpha Y + \beta] = \alpha \mathbb{E}[Y] + \beta$

"Expected value is linear"

Variance

If Y is a random variable, its variance is defined by

$$\mathrm{Var}[\underset{\substack{\uparrow \\ \text{random} \\ \text{variable}}}{Y}] = \mathbb{E}\Big[\big(Y - \underset{\substack{\uparrow \\ \text{Not} \\ \text{random}}}{\mathbb{E}[Y]}\big)^2\Big] = \mathbb{E}[\underset{\substack{\uparrow \\ \text{random} \\ \text{variable}}}{Y^2}] - \mathbb{E}[Y]^2$$

$\mathrm{Var}[Y]$ is a deterministic quantity (does not depend on the value of Y but depends on its distribution)

Properties:

- $Y \sim \{p_i\} \Rightarrow \mathrm{Var}[Y] = \sum_i p_i y_i^2 - \left(\sum_i p_i y_i\right)^2$
- $Y \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \mathrm{Var}[Y] = \sigma^2$
- $\forall (\alpha,\beta) \in \mathbb{R}^2, \; \mathrm{Var}[\alpha Y + \beta] = \alpha^2 \mathrm{Var}[Y]$

Remark: The most common densities (for continuous random variables) are log-concave functions

Ex) $$p(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

$$-\ln p(y) = + \ln\left(\sqrt{2\pi}\,\sigma\right) + \frac{1}{2\sigma^2}(y-\mu)^2$$ convex in y

② Markov inequality

Theorem: . Let Y be a nonnegative random variable ($\mathcal{Y} = [0, \infty)$) such that $\mathbb{E}[Y] < \infty$.

Then $\forall \varepsilon > 0$, $\boxed{\mathbb{P}(Y \geq \varepsilon) \leq \frac{\mathbb{E}[Y]}{\varepsilon}}$ Markov inequality

Other version: If $\mathcal{Y} = \mathbb{R}$ and $\mathbb{E}[|Y|] < \infty$, then $\forall \varepsilon > 0$,

$$\mathbb{P}(|Y| \geq \varepsilon) \leq \frac{\mathbb{E}[|Y|]}{\varepsilon}$$

↳ Basis for many concentration inequalities

→ With an inequality like Markov, can determine if a random variable concentrates around an interval or around its mean

→ Corollary of Markov: Chebyshev inequality

If Y is a random variable such that $\mathrm{Var}[Y] < \infty$,

Then $\mathbb{P}(\underbrace{|Y - \mathbb{E}[Y]|}_{\text{Distance to the mean}} \geq \varepsilon) \leq \frac{\mathrm{Var}[Y]}{\varepsilon^2}$

↳ Under the theorem's assumptions, Markov inequality is sharp (i.e. is the best inequality we can prove)

Proof based on convex optimization

Goal: Given Y nonnegative random variable, with $\mathbb{E}[Y] < \infty$ find a bound on $\mathbb{P}(Y \geq \varepsilon)$.

→ Given what we know about Y, we seek a bound that depends on $\mathbb{E}[Y]$

⇒ Simplest possible form: $\alpha \mathbb{E}[Y] + \beta$ for some $(\alpha, \beta) \in \mathbb{R}^2$.

⇒ We want to find α and β such that

$$\alpha \mathbb{E}[Y] + \beta \geq \mathbb{P}(Y \geq \varepsilon)$$

It suffices to find α and β such that

$$(1) \begin{cases} \alpha y + \beta \geq 1 & \forall y \geq \varepsilon \\ \alpha y + \beta \geq 0 & \forall y \in [0, \varepsilon) \end{cases}$$

Indeed, if α, β satisfy (1), then

$$\underbrace{\int_{[0,\infty)} (\alpha y + \beta) p(y) dy}_{\mathbb{E}[\alpha Y + \beta] = \alpha \mathbb{E}[Y] + \beta} \geq \underbrace{\int_{[0,\infty)} g(y) p(y) dy}_{\int_{[\varepsilon,\infty)} p(y) dy = \mathbb{P}(Y \geq \varepsilon)}$$

where $g(y) = \begin{cases} 1 & \text{if } y \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$

$$\alpha y + \beta \geq 1 \qquad \forall y \geq \varepsilon$$

$$\alpha y + \beta \geq 0 \qquad \forall y \in [0, \varepsilon)$$

$$\Rightarrow \begin{array}{l} (\alpha y + \beta)\, p(y) \geq 1 \times p(y) \quad \forall y \geq \varepsilon \\ (\alpha y + \beta)\, p(y) \geq 0 \times p(y) \quad \forall y \in [0, \varepsilon) \end{array}$$

p : density of Y

Define $g(y) = \begin{cases} 1 & \text{if } y \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$, we have

$$(\alpha y + \beta)\, p(y) \geq g(y)\, p(y) \quad \forall y \in [0, \infty)$$

$$\Rightarrow \underbrace{\int_{[0,\infty)} (\alpha y + \beta)\, p(y)\, dy}_{\substack{= \mathbb{E}[\alpha Y + \beta] \\ = \alpha \mathbb{E}[Y] + \beta}} \geq \underbrace{\int_{[0,\infty)} g(y)\, p(y)\, dy}_{\substack{= \int_{[0,\varepsilon)} 0 \times p(y)\, dy \\ + \int_{[\varepsilon,\infty)} p(y)\, dy}}$$

$$= \int_{[\varepsilon,\infty)} p(y)\, dy = \mathbb{P}(Y \geq \varepsilon)$$

Finding the best possible bound corresponds to solving the optimization problem

$$\underset{(\alpha,\beta) \in \mathbb{R}^2}{\text{minimize}} \quad \alpha \mathbb{E}[Y] + \beta$$

$$\text{s.t.} \quad \begin{array}{ll} \alpha y + \beta \geq 1 & \forall y \geq \varepsilon \\ \alpha y + \beta \geq 0 & \forall y \in [0, \varepsilon) \end{array}$$

→ We can find the solution of this convex optimization problem explicitly

$$\alpha^* = \frac{1}{\varepsilon} \qquad \beta^* = 0$$

The set of all valid inequalities is given by the feasible set

$$\{(\alpha,\beta) \mid \alpha\varepsilon + \beta \geq 1,\ \beta \geq 0,\ \alpha \geq 0\}$$

But for any feasible (α,β),

$$\alpha\,\mathbb{E}[Y] + \beta \geq \alpha^*\mathbb{E}[Y] + \beta^* \geq \mathbb{P}(Y \geq \varepsilon)$$