

Math Review

Sets and notation: $\{0, 1, 2\}, \mathbb{N}, \mathbb{R}$

$\in, \subseteq, \not\subseteq, \subsetneq, \not\subset, \cup, \cap, \setminus, \text{set}^c$

Set builder notation:

$$\{x \in \mathbb{N} \mid x < 3\} = \{0, 1, 2\}$$

$$\{f \mid f: \mathbb{R} \rightarrow \mathbb{R}\}$$

Cartesian products (Set of tuples):

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$$

$$\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R} = \{(a, b, c) \mid a, b, c \in \mathbb{R}\}$$

Dot product: $\vec{x} \in \mathbb{R}^d, \vec{w} \in \mathbb{R}^d$

$$\begin{aligned}\vec{x}^T \vec{w} &= (x_1, \dots, x_d) (w_1, \dots, w_d)^T \\ &= x_1 w_1 + \dots + x_d w_d\end{aligned}$$

Functions:

$$f: \mathbb{R} \rightarrow \mathbb{R}, f(x) = x^2$$

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}, f(\vec{x}) = f(x_1, x_2) = 2x_1 + 3x_2^2$$

$$A: (\mathbb{R} \times \mathbb{R})^n \rightarrow \{f \mid f: \mathbb{R} \rightarrow \mathbb{R}\}, \quad n=2$$

$$D = ((1, 2), (3, 4))$$

$$A(D) = f \quad \text{where} \quad f(x) = (4-2)x + 20 + 1$$

Summation and Integration:

$$\mathcal{X} = (x_1, x_2, x_3), \quad f(x) = x^2$$

$$\sum_{x \in \mathcal{X}} x = \sum_{i=1}^3 x_i = x_1 + x_2 + x_3$$

$$\sum_{x \in \mathcal{X}} f(x) = \sum_{i=1}^3 f(x_i) = f(x_1) + f(x_2) + f(x_3)$$

$$\mathcal{Y} = [a, b], \quad f(y) = y^c$$

$$\int_{\mathcal{Y}} f(y) dy = \int_a^b f(y) dy = \int_a^b y^c = \frac{y^{c+1}}{c+1} \Big|_a^b$$

$$f(x, y) = xy, \quad \mathcal{X} = (x_1, x_2, x_3), \quad \mathcal{Y} = [a, b]$$

$$\int_{\mathcal{Y}} \sum_{x \in \mathcal{X}} f(x, y) dy = \int_a^b \sum_{i=1}^3 x_i y dy$$

$\int_a^b$

$$= \int_a^b x_1 y + x_2 y + x_3 y \, dy$$

$$= \frac{x_1 y^2}{2} \Big|_a^b + \frac{x_2 y^2}{2} \Big|_a^b + \frac{x_3 y^2}{2} \Big|_a^b$$

(Partial) Derivatives:

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) = x^2, \quad f'(x) = \frac{df}{dx}(x) = 2x, \quad f''(x) = 2$$

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}, \quad f(\vec{x}) = f(x_1, x_2) = x_1^2 + x_2^2$$

$$\frac{\partial f}{\partial x_1}(x_1) = 2x_1, \quad \frac{\partial f}{\partial x_2}(x_2) = 2x_2, \quad \nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right)$$

Common derivatives and properties
on formula sheet

Probability

Outcome space and Events:

$\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ outcome space

$\tilde{E} \subseteq \mathcal{X}$

event

$\mathcal{X} \subseteq \mathcal{X}$
event

Probability Distribution:

P takes as input events and
outputs values in $[0, 1]$

$P(\tilde{E})$ where $\tilde{E} \subseteq \mathcal{X}$ probability of event $\tilde{E}$

Random Variable:

$X \in \mathcal{X}$ and has distribution P

$P(X \in \tilde{E}) \stackrel{\text{def}}{=} P(\tilde{E})$ where $\tilde{E} \subseteq \mathcal{X}$

Common notation: If $\tilde{E}$ contains a single outcome

$\tilde{E} = \{x\}$ where $x \in \mathcal{X}$. Then $P(X=x) \stackrel{\text{def}}{=} P(X \in \{x\})$

If $\tilde{E}$ is an interval:

$E = [a, b]$ then $P(a \leq X \leq b) \stackrel{\text{def}}{=} P(X \in [a, b])$

$E = [a, \infty)$ then $P(X \geq a) \stackrel{\text{def}}{=} P(X \in [a, \infty))$

$E = (-\infty, b]$ then $P(X \leq b) \stackrel{\text{def}}{=} P(X \in (-\infty, b])$

Discrete r.v.:

Countable outcome space $\mathcal{X} = \{2, 3, 4, 5\}$

Continuous r.v.:

Uncountable outcome space $\mathcal{X} = \mathbb{R}$

Calculating Probabilities:

If X is discrete: use pmf $p: \mathcal{X} \rightarrow [0, 1]$

$$P(X \in E) \stackrel{\text{def}}{=} \sum_{x \in E} p(x) \quad \text{where } E \subseteq \mathcal{X}$$

If Y is continuous: use pdf $p: \mathcal{X} \rightarrow [0, \infty)$

$$P(X \in E) \stackrel{\text{def}}{=} \int_E p(x) dx \quad \text{where } E \subseteq \mathcal{X}$$

commonly used discrete and continuous distributions on formula sheet.

Ex: Bernoulli, Normal, Laplace

Multivariate Probability:

$$Z = (X, Y) \in \mathcal{X} \times \mathcal{Y} \quad \text{r.v.}$$

On formula sheet $\mathcal{E}_x \subseteq \mathcal{X}, \mathcal{E}_y \subseteq \mathcal{Y}$

Joint: $P(X \in \mathcal{E}_x, Y \in \mathcal{E}_y)$

Marginal: $P_x(X \in \mathcal{E}_x), \quad P_y(Y \in \mathcal{E}_y)$

Conditional: $P_{X|Y}(X \in \mathcal{E}_x | Y=y), \quad P_{Y|X}(Y \in \mathcal{E}_y | X=x)$

Chain Rule: $p(x, y) = p(y|x) p(x) = p(x|y) p(y)$

Independence:

$X = (X_1, \dots, X_n)$ $X_1, \dots, X_n$ are independent if:

$$p(x_1, \dots, x_n) = p_{X_1}(x_1) p_{X_2}(x_2) \dots p_{X_n}(x_n)$$

Functions of r.v.:

A function of a r.v. is a r.v.

If $X \in \mathcal{X}$ is a r.v. then:

$f: \mathcal{X} \rightarrow \mathcal{Y}$, $Y = f(X) = X^2$ is a r.v. with

$$\bar{X} = g(X_1, \dots, X_n) = \frac{X_1 + \dots + X_n}{n} \quad \text{outcome space } \mathcal{Y}$$

Expectation and Variance:

$$Z = (X, Y) \in \mathcal{X} \times \mathcal{Y} \quad \text{r.v.}$$

On formula sheet with useful properties

Univariate: $\mathbb{E}[X]$

function: $\mathbb{E}[f(X)]$

Multivariate: $\mathbb{E}[f(Z)] = \mathbb{E}[f(X, Y)]$

Conditional: $\mathbb{E}[f(Y) | X = x]$

Variance: $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Supervised Learning

Dataset:

$$D = ((\vec{X}_1, Y_1), \dots, (\vec{X}_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n \quad \text{where}$$

$(\vec{X}_i, Y_i) \sim P_{\vec{X}, Y}$ and independent for all $i \in \{1, \dots, n\}$

$\mathcal{X} = \mathbb{R}^d$ feature vector (always $\mathbb{R}^d$)

$\mathcal{Y}$ Label or target

Learner:

$$A: (\mathcal{X} \times \mathcal{Y})^n \rightarrow \{f \mid f: \mathcal{X} \rightarrow \mathcal{Y}\}$$

Predictor:

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

loss function

$$l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R} \quad l(f(\vec{X}), Y)$$

Expected loss

$$L(f) = \mathbb{E}[l(f(\vec{X}), Y)] \quad (\vec{X}, Y) \sim P_{\vec{X}, Y}$$

Objective:

Define A such that $\mathbb{E}[L(A(D))]$ is small

Regression:

If $\mathcal{Y}$ has a notion of order

Usually $\mathbb{R}$ or an interval $[a, b]$

Use squared or absolute loss

Classification:

$\mathcal{Y}$ does not have a notion of order

Usually finite set like $\{\text{cat}, \text{dog}, \text{bird}\}$

Use 0-1 loss

Learner: ERM input dataset D

Estimation: Use D to estimate $L(f)$ for all $f \in \mathcal{F}$
call estimate $\hat{L}(f)$

Optimization: pick $\hat{f}$ as the $f \in \mathcal{F}$ that
minimizes $\hat{L}(f)$

Estimation:

$X \in \mathcal{X}$ is a r.v. with distribution $\mathbb{P}$

Want to estimate $\mathbb{E}[X] = \mu$

Use n i.i.d. samples from $\mathbb{P}$

$(X_1, \dots, X_n)$

Sample mean estimate:

$$\hat{\mu} = \bar{X} = g(X_1, \dots, X_n) = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

Expectation and Variance:

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X] = \mathbb{E}[X_1] = \dots = \mathbb{E}[X_n]$$

$$\text{Var}[\bar{X}] = \frac{\text{Var}[X]}{n} = \frac{\text{Var}[X_1]}{n} = \dots = \frac{\text{Var}[X_n]}{n}$$

Optimization

$$\min_{w \in W} g(w) = g(w^*) \quad \text{where } w^* = \arg\min_{w \in W} g(w)$$

$$w^* = \arg\min_{w \in W} g(w) = \arg\max_{w \in W} -g(w)$$

$$\min_{w \in W} g(w) = - \left(\max_{w \in W} -g(w) \right)$$

Solving Optimization problems:

- If W discrete just compare $g(w)$ for all $w \in W$
- If W continuous can use derivatives sometimes

Continuous Optimization:

If $g(w)$ is convex and twice differentiable then:

Cases: 1. If $W = \mathbb{R}$ then w^* is the solution to $g'(w) = 0$

2. If $\mathcal{W}=[a,b]$ then w^* is the solution to $g'(w)=0$ if this solution is in $[a,b]$. Otherwise, w^* is a or b

Twice differentiable: The second derivative of $g(w)$ written $g''(w)$ exists for all $w \in \mathcal{W}$

Convex: $g(w)$ is convex if $g''(w) \geq 0$ for all $w \in \mathcal{W}$

"Usually $g(w)$ is bowl shaped"

Multidimensional Minimization

If $\mathcal{W} = \mathbb{R}^d$ for $d > 1$, and $g(\vec{w})$ is convex

Note: it is more complicated to check if $g(\vec{w})$ is convex if $d > 1$. So, I will just tell you

Then we calculate

$$\vec{w}^* = (w_1^*, \dots, w_d^*)^T = \arg\min_{\vec{w} \in \mathcal{W}} g(\vec{w})$$

by setting w_j^* as the solution to

$$\frac{\partial g}{\partial w_j}(w_j) = 0 \quad \text{for all } j \in \{1, \dots, d\}$$

$$g(\vec{w}^*) = \min_{\vec{w} \in \mathcal{W}} g(\vec{w})$$