Mathematical Foundations of Data Science
Formula Sheet

Sums and Series

i. $\displaystyle\sum_{i=1}^{n} 1 = 1 + \ldots + 1 = n$

ii. $\displaystyle\sum_{j=0}^{n} ar^j = a + ar + \ldots + ar^n = a\frac{1-r^{n+1}}{1-r}$

iii. $\displaystyle\sum_{i=1}^{n} i = 1 + 2 + \ldots + n = \frac{n(n+1)}{2}$

iv. $\displaystyle\sum_{i=1}^{n} i^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} = \frac{n}{6}(2n+1)(n+1)$

p-series: $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges if $p > 1$ and diverges if $p \leq 1$.

Geometric series: $\displaystyle\sum_{n=0}^{\infty} x^n$ converges if $|x| < 1$ and diverges if $|x| \geq 1$.

Vanishing criterion: If $\displaystyle\sum_{n=0}^{\infty} a_n$ converges, then $a_n \to 0$ as $n \to \infty$. If $\displaystyle\lim_{n\to\infty} a_n \neq 0$, the series diverges.

Permutation and Combinations

	Without repetition	With repetition
Ordered (permutation)	$P_r^n = \dfrac{n!}{(n-r)!}$	n^r
Unordered (combination)	$\dbinom{n}{r} = \dfrac{n!}{(n-r)!r!}$	$\dbinom{r+(n-1)}{r} = \dbinom{r+(n-1)}{n-1}$

Discrete Probability

Bayes' theorem: $P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$

Law of total probability: $P(A) = \displaystyle\sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$, where the B_i's form a partition of the sample space S.

Generalised Bayes' theorem: For events A and B_i, $i = 1, \ldots, n$: $P(B_i|A) = \dfrac{P(A|B_i)P(B_i)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)}$, where the B_i's form a partition of the sample space, S.

Expected value: $E[X] = \displaystyle\sum_i x_i P(X = x_i)$

Variance: $\mathrm{var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$. $\quad E[X^2] = \displaystyle\sum_i x_i^2 P(X = x_i)$.

Bernoulli distribution:	$P(X=1) = p$ $P(X=0) = (1-p)$	$E[X] = p$	$\mathrm{var}(X) = p(1-p)$
Binomial distribution:	$P(X=k) = \dbinom{n}{k} p^k (1-p)^{n-k}$	$E[X] = np$	$\mathrm{var}(X) = np(1-p)$
Poisson distribution:	$P(X=k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$	$E[X] = \lambda$	$\mathrm{var}(X) = \lambda$

Linear Regression

$$\boldsymbol{y} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix} = X\boldsymbol{\beta} + \boldsymbol{e}.$$

The optimal solution is given by $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$.

Principal Component Analysis

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} X_i j \qquad X' = X - \mathbf{1}_{n\times 1}\bar{\boldsymbol{x}} \qquad C = \frac{1}{n-1}(X')^T X'$$

Integration

Antiderivatives:

i. $\int x^n dx = \frac{x^{n+1}}{n+1} + c,\ n \neq -1$

ii. $\int \frac{1}{x} dx = \log(x) + c$

iii. $\int e^{\lambda x} dx = \frac{e^{\lambda x}}{\lambda} + c$

iv. $\int \cos(ax) dx = \frac{\sin(ax)}{a} + c$

v. $\int \sin(ax) dx = -\frac{\cos(ax)}{a} + c$

vi. $\int cf(x) dx = c\int f(x) dx$

vii. $\int f(x) + g(x) dx = \int f(x) dx + \int g(x) dx$

Continuous Probability

Expected value: $E[X] = \int_{-\infty}^{\infty} x f(x) dx$ **Variance:** $\mathrm{var}(X) = \int_{-\infty}^{\infty} x^2 f(x) dx - \left(\int_{-\infty}^{\infty} x f(x) dx\right)^2$

Uniform distribution: $f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$ $E[X] = \frac{a+b}{2}$ $\mathrm{var}(X) = \frac{(b-a)^2}{12}$

Normal distribution: $f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ $E[X] = \mu$ $\mathrm{var}(X) = \sigma^2$

Exponential distribution: $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$ $E[X] = \frac{1}{\lambda}$ $\mathrm{var}(X) = \frac{1}{\lambda^2}$

Taylor and Maclaurin Polynomials

Taylor polynomial of degree n: $P_n(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!} + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n$.

Remainder term: $R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-a)^{n+1}$ for some z between a and x.

Important Taylor series:

i. $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$

ii. $\sin(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$

iii. $\cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$

iv. $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$

Gradient Descent

To (approximately) find the minimum of $f(m)$, the step size is given by $h = -\eta f'(m)$, for some learning rate η.