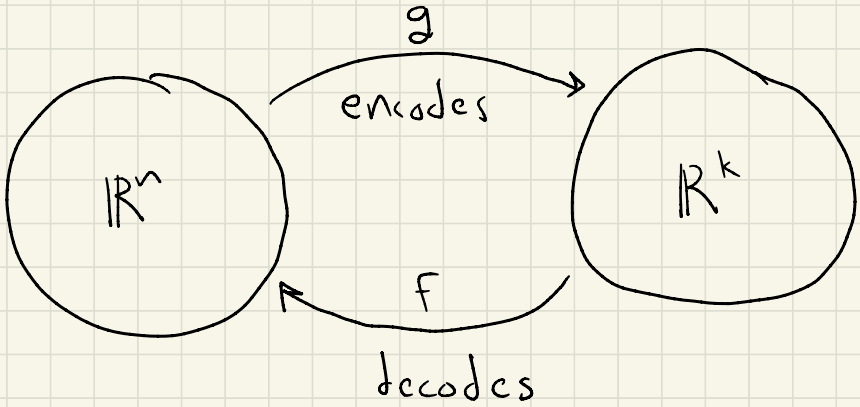
MAT 180 Notes

Principal Component Analysis

Dataset $\{x^{(1)}, \ldots, x^{(m)}\} \in \mathbb{R}^n$.

Goal: compress data to $\mathbb{R}^k$

Want functions



s.t. $x \approx f(g(x))$

Fix $f(y) = Dy$ $\quad D \in \mathbb{R}^{n \times k}$

D orthonormal columns

(i.e. $D^T D = I$)

To be optimal, we want

$$g(x) = \underset{y}{\operatorname{argmin}} \|x - f(y)\|_2^2$$

$$= \operatorname{argmin}_y (x-f(y))^T (x-f(y))$$

$$= \operatorname{argmin}_y \left(x^T x - f(y)^T x - x^T f(y) + f(y)^T f(y)\right)$$

$$= \operatorname{argmin}_y \left(-2x^T f(y) + f(y)^T f(y)\right)$$

$$= \operatorname{argmin}_y \left(-2x^T Dy + y^T D^T Dy\right)$$

$$= \operatorname{argmin}_y \left(-2x^T Dy + y^T y\right)$$

argmin means the argument y that minimizes

Solve the optimization problem

$$\nabla_y \left(-2x^T Dy + y^T y\right) = 0$$

Rule: $\nabla_y (My) = y^T \nabla_y (M^T) + M \cdot \nabla_y$

$$\Rightarrow -2x^T D + y^T \nabla(y) + y^T \nabla(y) = 0$$

$$\Rightarrow -2x^T D + 2y^T = 0 \Rightarrow y^T = x^T D$$

$$\Rightarrow \boxed{y = D^T x}$$

The optimal solution is

$$g(x) = D^T x.$$

★ still need D

Recall: we want $f(g(x))$ as close as possible to x for all points in the data set. That is,

$$D = \underset{C : C^T C = I}{\operatorname{argmin}} \sum_{i=1}^{m} \left\| x^{(i)} - \underbrace{CC^T x^{(i)}}_{f(g(x^{(i)}))} \right\|_2^2$$

$$= \underset{}{\operatorname{argmin}} \left\| X - XCC^T \right\|_F^2$$

$$X = \begin{pmatrix} - x^{(1)} - \\ \vdots \\ - x^{(m)} - \end{pmatrix} \in \mathbb{R}^{m \times n}$$

$$= \| \operatorname{Tr}\left((X - XCC^T)^T (X - XCC^T)\right)$$

$$= \| \operatorname{Tr}\left(X^TX - CC^TX^TX - X^TXCC^T + CC^TX^TXCC^T\right)$$

$$= \| \left(-\operatorname{Tr}(CC^TX^TX) - \operatorname{Tr}(X^TXCC^T) + \operatorname{Tr}(CC^TX^TXCC^T)\right)$$

$$= \| \left(-2\operatorname{Tr}(CC^TX^TX) + \operatorname{Tr}(X^TXCC^T)\right)$$

$$= \underset{C : C^TC = I}{\operatorname{argmin}} \; -\operatorname{Tr}(X^TXCC^T)$$

$$= \underset{C : C^TC = I}{\operatorname{argmax}} \operatorname{Tr}(C^TX^TXC)$$

Lemma The optimal encoding function is $g(x) = D^Tx$ where

$$D = \underset{C : C^TC = I}{\operatorname{argmax}} \operatorname{Tr}(C^TX^TXC)$$

To proceed, consider $k=1$, so C is a unit vector.

Goal: maximize $\mathrm{Tr}(C^T X^T X C) = C^T \underbrace{X^T X} C$

Note: $X^T X$ is positive semidefinite, symmetric

$\Rightarrow$ $X^T X$ is diagonalizable with orthonormal eigenvectors $v_1, \ldots, v_n$

eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$

Spectral Theorem

So $C \in \mathbb{R}^n \Rightarrow C = \sum_i a_i v_i \quad \sum a_i^2 = 1$

$$\Rightarrow C^T X^T X C = \left(\sum_i a_i v_i\right)^T X^T X \left(\sum_j a_j v_j\right)$$

$X^T X v_j = \lambda_j v_j$

$v_i^T v_j = \begin{cases} 1 & i=j \\ 0 & \text{else} \end{cases}$

$$= \sum_{i,j} a_i a_j v_i^T \lambda_j v_j$$

$$= \sum_i a_i^2 \lambda_i$$

largest value is λ_1

maximized when $a_1 = 1 \quad a_2 = \ldots = a_n = 0$

So $C = V_1$ (First principal component)

the eigenvector of $X^T X$ corresponding to the largest eigenvalue

Claim In general, the optimal solution is

$$D = C_{min} = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_k \\ | & & | \end{bmatrix}$$

eigenvectors of $X^T X$ corresponding to largest k eigenvalues

principal components of X

Proof: (induction) base case above (HW)

§3.1 Probability & Information Theory

9/28/22

(non-rigorous) (for rigour need Measure theory.)

§3.2

A random variable (RV) is a variable that takes on values from a specified set Ω of "states"

E.g. Outcomes of some random experiment

Notation X RV

x_1, x_2, x_3 values that X takes

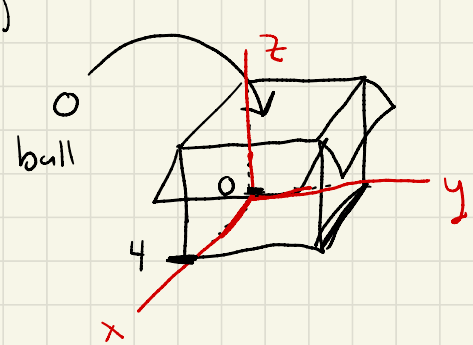
(discrete)

E.g. X measures outcomes of rolling a 6-sided dic. Maybe $x_1 = 2$ $x_2 = 6$ $x_3 = 1$, ...

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

(continuous)

E.g.



X = resting x-coord

$\Omega = [0, 4]$

$x_1 = 1.213$

$x_2 = \pi$

§3.3 Probability Distributions

How likely is each state of a RV?

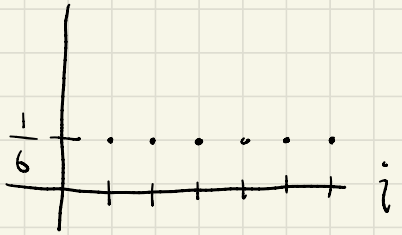
Discrete Probability Mass Function

$$P: \Omega \to [0,1] \quad \text{s.t.} \quad \sum_{x \in \Omega} P(x) = 1$$

Notation $P(x) = P(X = x)$ probability of x occurring

E.g. x = 6-sided die roll

$i \in \{1, \ldots, 6\}$, $P(i) = \frac{1}{6}$ $\qquad \sum_{i=1}^{6} P(i) = 1.$



E.g. x = two die rolls (6-sided, 8-sided)

$$\Omega = \{(i,j) \mid 1 \le i \le 6,\ 1 \le j \le 8\}$$

$$P(i,j) = \frac{1}{48}$$

Joint distribution

X = outcome of 6-sided die

Y = " " 8- " "

$\Omega_X = \{1, \dots, 6\}$ $\quad$ $\Omega_Y = \{1, \dots, 8\}$

$P_X(i) = \frac{1}{6}$ $\quad$ $P_Y(j) = \frac{1}{8}$

$$P(X=i, Y=j) = \frac{1}{48} = \frac{1}{6} \cdot \frac{1}{8} = P_X(i) P_Y(j)$$

↑ independent variables

E.g. Two 6-sided die rolls (indistinguishable)

$$\Omega = \{ \{i,j\} \mid 1 \le i,j \le 6 \}$$

$$P(\{i,i\}) = \frac{1}{36} \qquad 1 \le i \le 6$$

12 13 14 15 16
23 24 25 26
34 35 36
45 46
56

$$P(\{i,j\}) = \frac{2}{36} \qquad 1 \le i \ne j \le 6$$

$$6 \cdot \frac{1}{36} + 15 \cdot \frac{2}{36} = \frac{6+30}{36} = 1$$

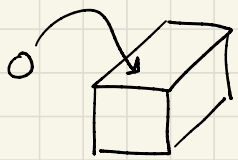
Continuous Ω uncountable ($\mathbb{R}$ or $[0,1]$)

probability density function

$$p : \Omega \to \mathbb{R}_{\geq 0} \quad \text{s.t.} \quad \underbrace{\int_\Omega p = 1}_{\text{Lebesgue integral}}$$

Interpretation: the probability that the outcome is in $S \subseteq \Omega$ is $\int_S p$

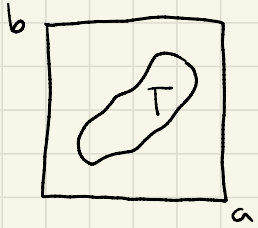
E.g.



$$\Omega = \underbrace{[0,a]}_{x} \times \underbrace{[0,b]}_{y}$$

$$p(x = x, y = y) = c$$

$$\int_\Omega p = 1 \Longrightarrow \int_0^a \int_0^b p \, dy \, dx = cab \Longrightarrow c = \frac{1}{ab}$$



$$P(T) = \int_T p = \iint_T \frac{1}{ab} \, dA = \frac{\text{area}(T)}{ab}$$

§ 3.4 Marginal Probability

Joint dist on x, y

$$p(x = x) = \sum_y P(X = x, Y = y)$$ discrete

$$p(x = x) = \int p(x = x, y = y)\, dy$$ continuous

---

§ 3.5 Conditional Probability

Probability that $X = x$ given that $Y = y$

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

E.g 2 - 6 sided die X = outcome $\{i, j\}$

Y = value of $i + j$

$$p(X = \{3,4\} \mid Y = 7) = \frac{p(X = \{3,4\}, Y = 7)}{p(Y = 7)}$$

$$= \frac{p(X = \{3,4\})}{p(Y = 7)} = \frac{\frac{2}{36}}{\frac{1}{6}} = \boxed{\frac{1}{3}}$$

Chain Rule

Bayes Rule

Independence of RV

} read these

Expectation value of $f(x)$ w.r.t dist P

$f: \Omega \to \mathbb{R}$

Expectation value: most likely outcome $f(x)$

$$E_P[f] = \sum_{x \in \Omega} p(x) f(x) \qquad \text{discrete}$$

$$= \int_\Omega pf \qquad \text{continuous}$$