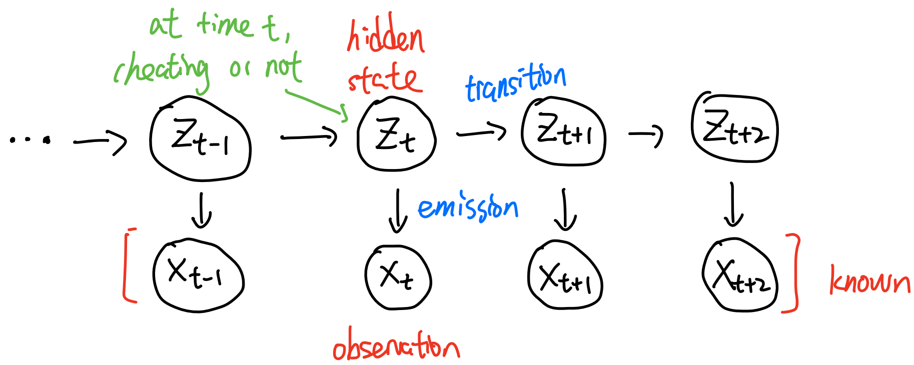
Inference on z_t



$$P(z_t \mid x_{1:T}) \propto p(z_t, x_{1:T}) = \underbrace{P(x_{1:t}, z_t)}_{\substack{\alpha_t(z_t) \\ \text{forward}}} \underbrace{P(x_{t+1:T} \mid z_t)}_{\substack{\beta_t(z_t) \\ \text{backward}}}$$

α-recursion : suppose we know $\alpha_{t-1}(i) = P(x_{1:t-1}, z_{t-1} = i)$
for every i

What is $\alpha_t(j)$ for all j marginalise out z_{t-1}

$$\alpha_t(j) = \sum_{i=1}^{k} P(x_{1:t-1}, z_{t-1} = i) P(z_t = j \mid z_{t-1} = i) P(x_t \mid z_t = j)$$

$$= \sum_{i=1}^{k} \alpha_{t-1}(i) \cdot A_{ij} \cdot P(x_t \mid z_t = j)$$

Base case: $\alpha_1(j) = p(x_1, z_1 = j) = \underbrace{P(z_1 = j)}_{\text{prior}} \underbrace{P(x_1 \mid z_1 = j)}_{\text{emission}}$

For $\beta_t(j)$: compute it based on $\beta_{t+1}(i)$ for every i

Summary : To infer $P(z_t \mid x_{1:T})$

① compute α_t's and β_t's recursively

② compute $P(z_t = j \mid x_{1:T}) = \dfrac{\alpha_t(j)\, \beta_t(j)}{\sum_{i=1}^{k} \alpha_t(i)\, \beta_t(i)}$

Learning HMM parameters (EM)

idea of EM: if we had complete data, live is easier

z :	2	1	3	1	2	sequence 1
x :	1.6	6.1	2.2	9.2	1.4	

z :	1	3	2	1	sequence 2
x :	7.4	2.1	1.2	9.7	

$$P(z_1) = \begin{cases} \frac{1}{2} & z_1 = 1 \\ \frac{1}{2} & z_1 = 2 \\ 0 & z_1 = 3 \end{cases}$$

$P(x_t \mid z_t = 1) = N(8, 6)$

$P(x_t \mid z_t = 2) = N(1.4, 0.1)$

$$P(z_t \mid z_{t-1} = 1) = \begin{cases} 0 & z_t = 1 \\ 1/3 & z_t = 2 \\ 2/3 & z_t = 3 \end{cases}$$

Suppose we don't know any z_t's

E-step creates fictitious data

M-step estimates params on fictitious data

$P(z_1)$:

E-step: Compute $P(z_1 \mid x_{1:T})$

suppose get: sequence 1 : [0.2, 0.7, 0.1]
sequence 2 : [0.6, 0.3, 0.1]

pretend our data

10 copies of sequence 1 where $\begin{cases} 2 \text{ have } z_1 = 1 \\ 7 \text{ have } z_1 = 2 \\ 1 \text{ have } z_1 = 3 \end{cases}$

10 copies of sequence 2 where $\begin{cases} 6 \quad .. \quad z_1 \\ 3 \quad .. \quad z_2 \\ 1 \quad .. \quad z_3 \end{cases}$

M-step (estimate params)

treat fictitious data as real,
count things to estimate params

$P(z_1=1) = 8/20$

$P(z_1=2) = 10/20$

$P(z_1=3) = 2/20$

For emissions:

E-step: infer $P(z_t \mid x_{1:T})$ for every t

suppose for some t, $x_t = 1.7$

$P(z_t \mid x_{1:T}) = [0.7, \ 0.1, \ 0.2]$

we have 0.7 counts of $(z_1=1, x_t=1.7)$

0.1 $(z_1=2, x_t=1.7)$

0.2 $(z_1=3, x_t=1.7)$

Transitions:

E-step: we want pseudo-counts of how many times state $i \to$ state j

$P(z_{t-1}, z_t \mid x_{1:T})$

Based on observations, which pairs (z_{t-1}, z_t) are likely?

$$\begin{tikzcd}
\cdots & z_{t-1} \arrow[r] \arrow[d] & \boxed{z_t} \arrow[r] \arrow[d] & z_{t+1} \arrow[d] \\
& x_{t-1} & \boxed{x_t} & \boxed{x_{t+1}}
\end{tikzcd}$$

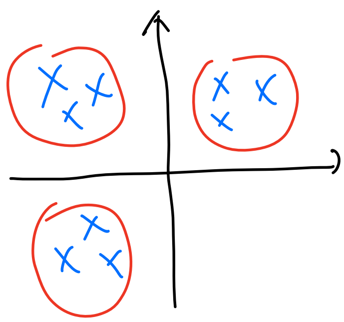
$$P(z_{t-1}, z_t, x_{1:T}) = \underbrace{P(x_{1:t-1}, z_{t-1})}_{=\,\alpha_{t-1}(z_{t-1})} \underbrace{P(x_{t+1:T} \mid z_t)}_{=\,\beta_t(z_t)}$$

$$\cdot\, P(z_t \mid z_{t-1})\, P(x_t \mid z_t)$$

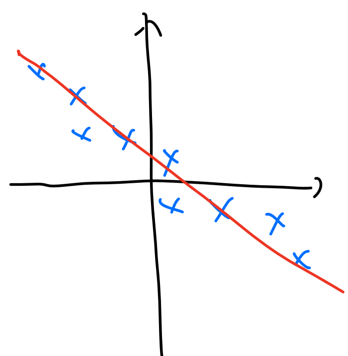
then normalise over all pairs of (z_{t-1}, z_t)

M-step: use these pseudocounts to estimate transition probabilities.

Dimensionality Reduction



clustering



dimensionality reduction

Given $\{x^{(1)}, \dots, x^{(n)}\} \in \mathbb{R}^d$

find a lower dimensional subspace

that preserve most of the information

Method: principle component analysis (PCA)

want to find a good 1-D projection

key assumption: data has mean 0

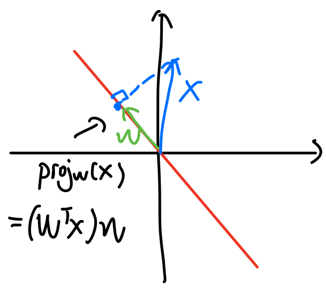
↳ in practice, compute mean of data, subtract

Parameter $w \in \mathbb{R}^d$ that defines 1-D subspace, $\|w\| = 1$

What loss function describes good choice of w?

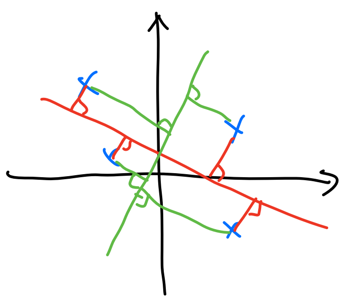
Reconstruction error

$$\sum_{i=1}^{n} \|x^{(i)} - \underbrace{\mathrm{Proj}_w(x^{(i)})}_{\text{projection of } x^{(i)} \text{ on } w}\|^2$$



$$= \sum_{i=1}^{n} \| x^{(i)} - (w^T x^{(i)}) \cdot w \|^2$$

PCA chooses w to minimize this loss function



large error

small error

equivalently: maximize variance of points after projection

by pythagorean theorem:

$$\underbrace{(w^T x)^2}_{\text{maximize}} + \underbrace{\text{ReconError}}_{\Leftrightarrow \text{ minimize}} = \underbrace{\|x\|^2}_{\text{fixed}}$$

$$\text{maximize} \sum_{i=1}^{n} (w^T x^{(i)})^2$$

$\frac{1}{n} \sum_{i=1}^{n} (w^T x^{(i)})^2$ is variance of $w^T x$