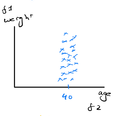
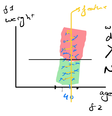
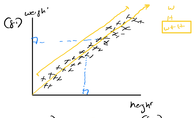
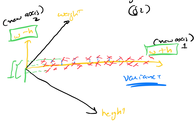
Principal Component Analysis

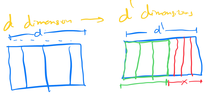
People with diabetes (y=1) / (y=0) — Diabetes









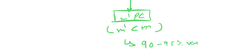


(100% var) n features $\xrightarrow{PCA}$ (100% var) m PC

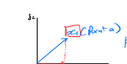
$\boxed{n \text{ PC}}$ $(n' < m)$

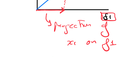
$\Rightarrow$ 90 - 95% var

$\text{Proj}_{d_1} a = \vec{a} \cdot \vec{d_1}$ (new basis) $(\vec{d_1} = 1)$



PCA → info → high var.





$\uparrow\uparrow$ proj $\Leftrightarrow$ $\perp\perp$

computation

$\mu_y = 0$

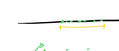
Σ and

$\mu_{d_1} = 0$

$\mu_{d_2} = 0$

$\vdots$

$\mu_{d_d} = 0$



$w^T x + w_0 = 0$

$w_0 = 0$

$\boxed{w^T x = 0}$

$\vec{w}^T \vec{x} = 0$

for any point x_i, projection on $\vec{u}$

$= \dfrac{\vec{x_i} \cdot \vec{u}}{\|u\|}$

$$\text{obj} = \max \sum_{i=1}^{n} \frac{\vec{x_i} \cdot \vec{u}}{\|\vec{u}\|}$$

assume (cond.) $\Rightarrow \|\vec{u}\| = 1$

$$\text{obj} = \frac{1}{n} \max \sum_{i=1}^{n} \vec{x_i} \cdot \vec{u}$$

st $\|\vec{u}\| = 1$

Lagrange multiplier

$\|\vec{u}\| = 1$

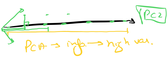
$g(x) = 0$

$= \|\vec{u}\| - 1 = 0$

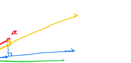
$$\text{obj} = \frac{1}{n} \max \sum_{i=1}^{n} (\vec{x_i} \cdot \vec{u}) + \lambda(\|\vec{u}\| - 1)$$

Projection should not cancel each other out

$$\frac{1}{n} \sum (\vec{x} \cdot \vec{u})^2 + \lambda(\|\vec{u}\| - 1)$$



$\underset{(n \times d)}{X} \; \underset{(d \times 1)}{\vec{u}} = \underset{(n \times 1)}{}$



$$\frac{1}{n} \sum (\underbrace{\vec{x} \cdot \vec{u}}_{\text{scalar data}})^2 + \lambda(\|\vec{u}\| - 1)$$

$$\max_{\vec{u}, \lambda} \frac{(\vec{X} \cdot \vec{u})^2}{n} + \lambda(\|\vec{u}\| - 1)$$

using $A^T A = A^2$

$$\max \frac{(X \cdot u)^T (X \cdot u)}{n} + \lambda(u^T u - 1)$$

$(X \cdot u)^T = u^T \cdot X^T$

$$\frac{u^T \cdot X^T \cdot X \cdot u}{n} + \lambda(u^T u - 1)$$

$$\mathcal{L} = u^T V u + \lambda(u^T u - 1)$$

$= u^T V u + \lambda u^T u - \lambda$

$$\frac{\partial \mathcal{L}}{\partial u} = 2Vu + 2\lambda u$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = u^T u - 1$$

$u^T u - 1 = 0$

$u^T u = 1$

$\|u\| = 1$

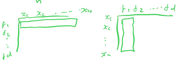
$$2Vu + 2\lambda u = 0$$

$$\underbrace{V}_{\text{covariance matrix}} \underbrace{\vec{u}}_{\text{eigen vector}} = \underbrace{-\lambda}_{\text{scalar constant}} \vec{u}$$

$\vec{u}$, $\vec{v}$, $\vec{w}$ — e_1, e_2, e_3 — (not (dependency) perfectly ind)

eigen value decomposition

$\dfrac{X^T X}{n}$



$\boxed{x_1\ x_2\ x_3 \cdots x_n}$ $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$$= \frac{x_1^2 + x_2^2 \cdots x_n^2}{n}$$

$$= \frac{\sum x_i^2}{n} \qquad \bar{x} = 0 \quad \boxed{\mu = 0}$$

$$= \sum \frac{(x_i - \bar{x})^2}{n} \quad \text{std var}$$

$e_1 > e_2 \cdots\cdots e_d$

Choose eigen values y more imp PC

$e_1 + e_2 \cdots e_d$ } % var retained

$e_1 + e_2 \cdots e_d$