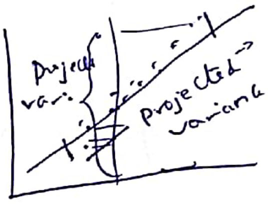
Principal Component Analysis.

Objective: Find a linear (orthogonal) Projection of the data such that the Projected variance is maximized.



$D = \{x_1, x_2 \ldots, x_n\}$

$\underline{x}_i \in R^d$.

Let $X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}_{d \times n}$ matrix constructed from the data - Data matrix.

Let us start by projecting the data onto a 1D manifold [line].

Let $u_1 \in R^d$ be the line on to which we are seeking the projection.

Let z represent the projected points.

we have

$$\underset{1\times n}{z_1} = \underset{1\times d}{u_1^T} \underset{d\times n}{X}$$

Every component of z corresponds to the projection of X on u_1.

Objective of PCA

$$u_1^* = \underset{u_1}{\arg\max} \; var(u_1^T X) \quad -(1)$$

$$var(u_1^T X) = u_1^T S u_1 \qquad \underset{d\times d}{S} = (X-\bar{X})(X-\bar{X})^T$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

observe that $u_1^T S u_1$ is a scalar & S is PSD & symmetric.

Thus (1) is ill-defined if it is unconstrained

However since we are only interested in the direction of projection we can constrain the problem by fixing the norm of u_1 to any constant.

Let $u_1^T u_1 = 1$.

Thus, $$u_1^* = \underset{u_1}{\arg\max} \left(u_1^T S u_1 \right)$$

$$\text{s.t } u_1^T u_1 = 1.$$

$$L = u_1^T S u_1 - \lambda \left(u_1^T u_1 - 1 \right)$$

$$\frac{\partial L}{\partial u_1} = 0 \Rightarrow S u_1 = \lambda u_1$$

$$\therefore u_1^T S u_1 = u_1^T \lambda u_1$$

$$= \underline{\lambda}.$$

Thus, $$u_1^* = \underset{u_1}{\arg\max} (\lambda).$$

But we have λ as the Eval of S.

Thus, u_1 will be the direction of the Evec of S corresponding to the maximum Eval of S.

Extending this, suppose the Evals of S are ordered in accordance to their value.

$$\begin{array}{ccccc} \lambda_1 > & \lambda_2 > & \lambda_3 > & \cdots & \lambda_d \\ | & | & & & \\ u_1 & u_2 & \cdots & \cdots & u_d \end{array}$$

Define

$$U = [u_1 \; u_2 \cdots u_d]_{d \times d}$$

$$\underset{d \times n}{Z} = \underset{d \times d}{U^T} \; \underset{d \times n}{X}$$ – represents the projection of X on to a new-co-ordinate system where the variance is maximized.

Since S is symmetric, U will be a orthonormal matrix – columns of U are called principal vectors.

Now if we know that the original data 'effectively lies' in a p-dimensional subspace of d [usually $p << d$]

Then one can consider

$$\underset{p\times n}{\hat{Z}} = \underset{p\times d}{\hat{U}^T}\underset{d\times n}{X}$$

where $\hat{U} = [\psi_1, \psi_2, \dots \psi_p]_{d\times p}$

Now $\hat{Z}$ will be a new set of datapoints lying in a p-dimensional subspace.

Thus, PCA can be used for dimensionality reduction.

$Z = U^T X$ since U is orthogonal,

$$UZ = UU^T X$$
$$= \underline{X}$$

Thus, one can recover back the original data from projections.

we can also reconstruct the data after we reduce the dimensions.

$$\hat{z} = \hat{U}^T x$$

Now $U\hat{z}$ (call it $\hat{x}$) will be the reconstruction of x based on the first P - principal componets.

Lets consider the second formulation.

$$\cancel{\hat{X} = \hat{U}\hat{z}}$$

$$\cancel{= \hat{U}\hat{U}^T x}$$

Error in the reconstruction is

$$e = \|x - \hat{x}\|_2 .$$

$$e = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 \quad - (1)$$

$$\hat{x}_i = \sum_{j=1}^{P} \alpha_{ij} u_j + \sum_{j=P+1}^{D} \beta_j u_j$$

$\hat{z}$ (imp thm)

substituting for $\hat{x}_i$ in (1),

$$\alpha_{ij} = x_i^T u_j \quad , \quad j = 1, \ldots P$$

$$\beta_j = \bar{x}^T u_j \qquad \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j ,$$

$j = P+1, \ldots D.$

Now, ~~$x_n - \hat{x}_n$~~ we have

$$x_n = \sum_{j=1}^{D} (x_i^T u_j) u_j$$

$$\therefore \quad x_i - \hat{x}_i = \sum_{j=P+1}^{D} \left[(x_i - \bar{x})^T u_j \right] u_j$$

Now, $e = \frac{1}{n} \sum_{j=P+1}^{D} u_j^T S u_j$ $\therefore$ minimizing e

$$e = \sum_{j=p+1}^{D} u_j^T S u_j$$

Similar to the previous case, e will be minimized when u_js are the ~~ta~~ Evecs of S corresponding to last D-P Evals..

Linear models.

$h(x)$ is of the form

$$h(x) = W^T \phi(x)$$

where ϕ is a fixed function of x.

ϕ can be polynomial, logistic sigmoid etc.

$h(x)$ is called linear because its a linear function in the parameter space W.

~~In all~~ We shall consider some of the families of these linear models.

Before we go to linear models, let us look at a general important result.

with our usual notations.

Let R denote the risk associated with a classifier $h(x)$.

Let us consider the squared error loss.

$$L(y, h(x)) = (h(x)-y)^2$$

$$R(h) = \iint L(y, h(x))\, P(x,y)\, dx\, dy$$

$$= \iint (h(x)-y)^2\, P(x,y)\, dx\, dy$$

Goal of ML: find h such that

$$h^* = \underset{h}{\operatorname{argmin}}\; R(h)$$

$$h^* = \operatorname{argmin} \iint (h(x)-y)^2\, P(x,y)\, dx\, dy.$$

$$\frac{\partial R}{\partial h} = 2\int (h(x) - y)\, p(x,y)\, dy$$

$$h(x) = \frac{\int y\, p(x,y)\, dy}{\int p(x,y)\, dy} = \frac{\int y\, p(x,y)\, dy}{p(x)}$$

$$= \int y\, p(y|x)\, dy = \underline{E_y[y|x]}$$

Thus, the optimal classifier is the conditional expectation of the labels given the data for squared error loss.

When $y \in \{0,1\}$, $E_y[y|x] = P[y=1|x] = \eta_1(x)$

Giving us back the Baye's classifier.

Now, lets come back to the linear models.

$$\cancel{y} \quad h(x) = W^T \phi(x).$$

Consider that ~~y~~ true y is being approximated by a deterministic function of x $h(x)$ with an uncertainity ϵ.

$$\therefore \quad y = h(x) + \epsilon$$
$$= W^T \phi(x) + \epsilon$$

Let us assume that $\epsilon \sim N(0, \sigma^2)$.

Now since we know that the optimal estimator is the conditional expectation,

we need $\quad E[y/x] = \int y \, f(y|x) \, dy$