

# MACHINE LEARNING

DEVANSH TRIPATHI

ABSTRACT. We shall make some short notes from machine learning textbook by Hui Jiang.

**Discriminative models:** They simply assume that the input samples and their corresponding output label are generated by an unknown function. These models attempt to estimate that function. It can be linear/bilinear/quadratic functions/neural networks (as universal function approximators).

**Generative models:** They assume both the input variable  $x$  and output variable  $y$  are random variables and they try to figure out their joint probability distribution from the data.

## 1. DIMENSIONALITY REDUCTION

**1.1. Linear Dimensionality Reduction.** PCA aims to search for some orthogonal projection directions in the space that can achieve the **maximum variance**. These directions are often called the **principal components** of the original data distribution.

These principal components are used as basis vectors to construct the linear subspace for dimensionality reduction.

The result in figure 2 shows that if we want to maximize the variance, we need to take the eigenvector corresponding to the **maximum eigenvalue**.

This result can be extended to the case where we want to map  $x \in \mathbb{R}^n$  into a lower dimensional space  $\mathbb{R}^m (m \ll n)$ . We need to take  $m$  eigenvectors corresponding to top  $m$  eigenvalues of the covariance matrix. These  $m$  eigenvectors are denoted as  $\{\hat{w}_1, \hat{w}_2, \dots, \hat{w}_m\}$  then the matrix  $A$  for transformation can be written as:

$$A = \begin{bmatrix} - & \hat{w}_1^T & - \\ - & \hat{w}_2^T & - \\ \vdots & & \\ - & \hat{w}_m^T & - \end{bmatrix}_{m \times n}$$

1

Projection of  $x$  in the direction of  $w \rightarrow$

$$v = x \cdot w = w^T x$$

$$\left\{ \begin{array}{l} v = \|x\| \cos \theta \text{ and} \\ \cos \theta = \frac{x \cdot w}{\|x\| \|w\|} \end{array} \right.$$

Assume we have set of  $N$  vectors  
in a  $n$ -dimensional space -

[ $\because \|w\| = 1$  assumption]

$$\mathcal{D} = \{x_1, x_2, \dots, x_N\}$$

And let  $\{v_1, v_2, \dots, v_N\}$  be their projection in the  
direction of  $w$ , where -

$$v_i = w^T x_i \quad \forall i$$

Variance of projection vectors -

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (v_i - \bar{v})^2 & \left[ \bar{v} = \frac{1}{N} \sum_{i=1}^N v_i \right] \\ &= \frac{1}{N} \sum_{i=1}^N (v_i - w^T \bar{x})^2 & = \frac{1}{N} \sum_{i=1}^N w^T x_i \\ & & = w^T \bar{x} \end{aligned}$$

$$\begin{aligned} \Rightarrow \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (v_i - w^T \bar{x}) (v_i - w^T \bar{x}) \\ &= \frac{1}{N} \sum_{i=1}^N (w^T x_i - w^T \bar{x}) (w^T x_i - w^T \bar{x}) \\ &= \frac{1}{N} \sum_{i=1}^N w^T (x_i - \bar{x}) w^T (x_i - \bar{x}) \\ &= \frac{1}{N} \sum_{i=1}^N w^T (x_i - \bar{x}) (x_i - \bar{x})^T w \\ \sigma^2 &= w^T \left[ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) (x_i - \bar{x})^T \right] w \quad \text{--- (1)} \end{aligned}$$

sample covariance matrix of  $\mathcal{D}$ .

Principal component  $\hat{w} = \arg \max_w w^T S w$  subject to  $w^T w = 1$

Lagrangian is as follows  $\rightarrow$

$$L(w) = w^T S w + \lambda \cdot (1 - w^T w)$$

FIGURE 1. Maximizing the variance

$$\frac{\partial \mathcal{L}(w)}{\partial w} = 2Sw - 2\lambda w = 0$$

$$S\hat{w} = \lambda \hat{w}$$

→ Principal component must be an eigenvector of  $S$  (sample covariance matrix) and  $\lambda$  is an eigenvalue

From eq<sup>n</sup> ① →  $\sigma^2 = \hat{w}^T S \hat{w} = \hat{w}^T \lambda \hat{w} = \lambda \cdot \|\hat{w}\|^2$

(Projection variance) →  $\sigma^2 = \lambda$

FIGURE 2. Maximizing the variance

Each eigenvector forms a row of  $A$ . Since, **the covariance matrix of  $S$  in figure 1 is always symmetric and has full rank.** Therefore, we can compute  $n$  different mutually orthogonal eigenvector for  $S$ .

**Covariance Matrix.** It is a generalization of the variance in the higher dimensions. Let  $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  be a random variable then

$$\text{var}(X) = \text{cov}(X, X) = E[(X - E[X])(X - E[X])^T]$$

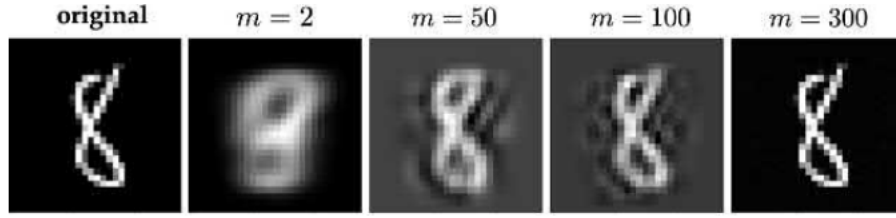
The diagonal entries corresponds to variance ( $\text{cov}[X_i, X_i] = \text{var}(X_i)$ ) and other than diagonal entries are covariances.

**Covariance Matrix**

$$\begin{bmatrix} \text{Var}(x_1) & \dots & \text{Cov}(x_n, x_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}(x_n, x_1) & \dots & \text{Var}(x_n) \end{bmatrix}$$

FIGURE 3. Covariance Matrix

**Reconstruction of  $x$  from  $y$ .** Let  $y = Ax$  where  $A$  is the matrix of the transformation and  $y$  is the transformed data. When we transform  $n$  dimension data to  $m$  dimensional space and  $m \ll n$  then recovering  $n$  dimension vector is not possible. If  $m \approx n$  only then recovering is possible to some extent. This



where the original image of a handwritten digit is  $28 \times 28 = 784$  in size,

FIGURE 4. Recovered images from lower dimensional space

### PCA Procedure

Assume the training data are given as  $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ .

1. Compute the sample covariance matrix  $\mathbf{S}$  in Eq. (4.3).
2. Calculate the top  $m$  eigenvectors of  $\mathbf{S}$ .
3. Form  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with an eigenvector in a row.
4. For any  $\mathbf{x} \in \mathbb{R}^n$ , map it to  $\mathbf{y} \in \mathbb{R}^m$  as  $\mathbf{y} = \mathbf{A}\mathbf{x}$ .

FIGURE 5. Summary of PCA