# 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 constuct 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 $\{\hat{w}_1, \hat{w}_2 \ldots, \hat{w}_m\}$ then the matrix $A$ for transformation can be written as:

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

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

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

$$\left[\begin{array}{l} V = \|x\| \cos\theta \text{ and} \\ \cos\theta = \dfrac{x \cdot w}{\|x\| \, \|w\|} \end{array}\right.$$

$[\because \|w\| = 1 \text{ assumption}]$

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

$$\mathcal{D} = \{x_1 \; x_2 \; — \; x_N\}$$

And let $\{v_1, v_2 \; —— \; v_N\}$ be their projection in the direction of $w$, where –

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

Variance of projection vectors –

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (v_i - \bar{v})^2 \qquad \left[\bar{v} = \frac{1}{N} \sum_{i=1}^{N} v_i\right.$$

$$= \frac{1}{N} \sum_{i=1}^{N} (v_i - w^T \bar{x})^2 \qquad = \frac{1}{N} \sum_{i=1}^{N} w^T x_i$$

$$= w^T \bar{x}$$

$$\Rightarrow \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (v_i - w^T \bar{x})(v_i - w^T \bar{x})$$

$$= \frac{1}{N} \sum_{i=1}^{N} (w^T x_i - w^T \bar{x})(w^T x_i - w^T \bar{x})$$

$$= \frac{1}{N} \sum_{i=1}^{N} w^T (x_i - \bar{x}) \, w^T (x_i - \bar{x})$$

$$= \frac{1}{N} \sum_{i=1}^{N} w^T (x_i - \bar{x}) (x_i - \bar{x})^T w$$

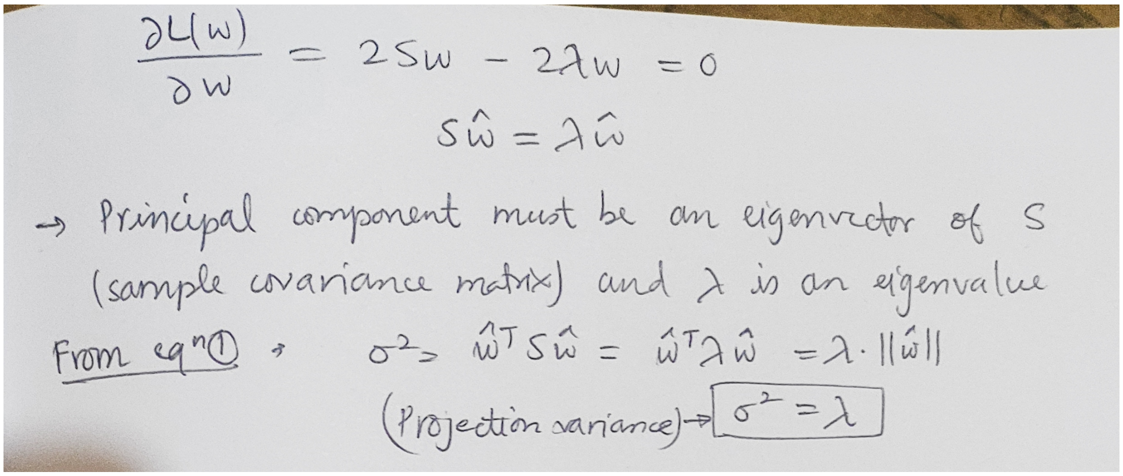
$$\sigma^2 = w^T \underbrace{\left[\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T\right]}_{\text{sample covariance matrix of } \mathcal{D}.} w \qquad —— (1)$$

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



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

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

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

From eq$^n$ ① : $\sigma^2 = \hat{w}^T S \hat{w} = \hat{w}^T \lambda \hat{w} = \lambda \cdot ||\hat{w}||$

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

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 $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}\}$ be a random variable then

$$var(X) = cov(X, X) = E[(X - E[X])(X - E[X])^T]$$

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

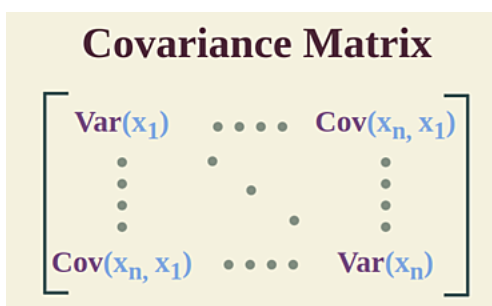



Figure 3. Covariance Matrix

**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

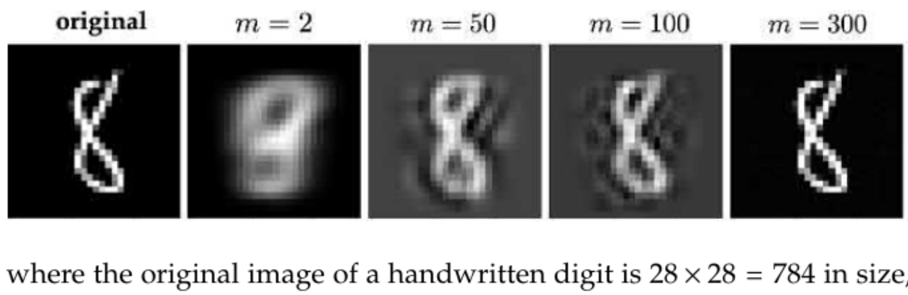



FIGURE 4. Recovered images from lower dimensional space

**PCA Procedure**

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

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