Machine Learning I: Fractal 2

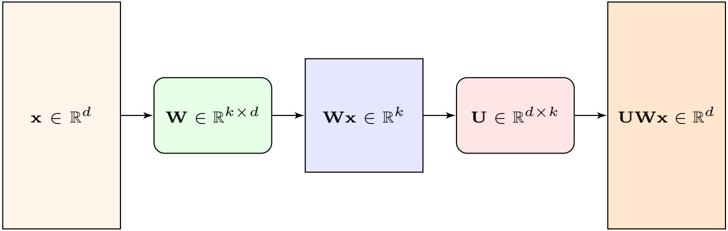
Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur

These slides are prepared from the following book:
Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

Dimensionality Reduction

- Let $\mathbf{X} = \left\{\mathbf{x}_i : \mathbf{x}_i \in \mathbb{R}^d\right\}_{i=1}^{n}$ be the input dataset.
- Let $\mathbf{W} \in \mathbb{R}^{k \times d}$ be a matrix, where $k < d$, then $\mathbf{W}\mathbf{x} \in \mathbb{R}^k$ is the lower dimensionality representation of $\mathbf{x}$.
- We can use a matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$ to recover the each original vector $\mathbf{x}$ from its compressed version.


$\mathbf{x} \in \mathbb{R}^d$
$\mathbf{W} \in \mathbb{R}^{k \times d}$
$\mathbf{W}\mathbf{x} \in \mathbb{R}^k$
$\mathbf{U} \in \mathbb{R}^{d \times k}$
$\mathbf{U}\mathbf{W}\mathbf{x} \in \mathbb{R}^d$

$$\underset{\mathbf{W}\in\mathbb{R}^{k\times d},\mathbf{U}\in\mathbb{R}^{d\times k}}{\text{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{x}_i\|_2^2.$$

Principal Component Analysis

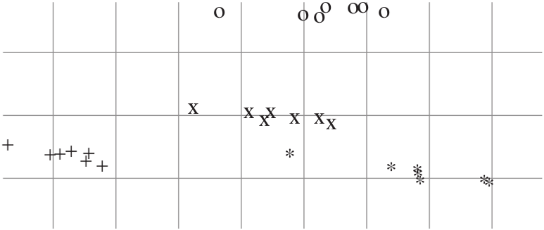
Problem

$$\max_{\mathbf{U}\in\mathbb{R}^{d\times k},\mathbf{U}^\top\mathbf{U}=\mathbf{I}} \text{Trace}(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U})$$

Solution

Let $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_n\end{bmatrix}$ and $\mathbf{X}\mathbf{X}^\top\mathbf{u}_i = \lambda_i\mathbf{u}_i, \forall i \in \{1, 2, \ldots, n\}$ be the EVD of the matrix $\mathbf{X}\mathbf{X}^\top$. Here, we assume that the eigenvalues are such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then, the solution to the above problem is $\mathbf{U}^\star = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k\end{bmatrix}$.

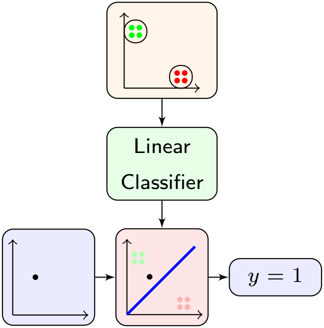
Face Recognition



pc-http://vision.ucsd.edu/content/yale-face-database

Generative Models

Let $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ be a training set. Then, our goal is to find a predictor h such that $h(\mathbf{x}_i)$ is equal to the true label of the input $\mathbf{x}_i$.



We do not impose any assumptions on the underlying distribution over the data $\mathcal{S}$. Our goal is not to learn the underlying distribution but rather to learn an accurate predictor

Generative Approach

- We describe a generative approach, in which it is assumed that the underlying distribution over the data has a specific parametric form and our goal is to estimate the parameters of the model. This task is called parametric density estimation.
- The discriminative approach has the advantage of directly optimizing the quantity of interest (the prediction accuracy) instead of learning the underlying distribution.
- The problem is that it is usually more difficult to learn the underlying distribution than to learn an accurate predictor.
- However, in some situations, it is reasonable to adopt the generative learning approach.

Maximum Likelihood Estimator

- A drug company developed a new drug to treat some deadly disease.
- We would like to estimate the probability of survival when using the drug.
- To do so, the drug company sampled a training set of m people and gave them the drug.
- Let $\mathcal{S} = \{x_1, x_2, \ldots, x_m\}$ denote the training set, where for each i, $x_i = 1$ if the i^{th} person survived and $x_i = 0$ otherwise.
- We can model the underlying distribution using a single parameter, $\theta \in [0, 1]$, indicating the probability of survival.

Maximum Likelihood Estimator

- We now would like to estimate the parameter θ on the basis of the training set $\mathcal{S}$.
- A natural idea is to use the average number of 1's in $\mathcal{S}$ as an estimator. That is,

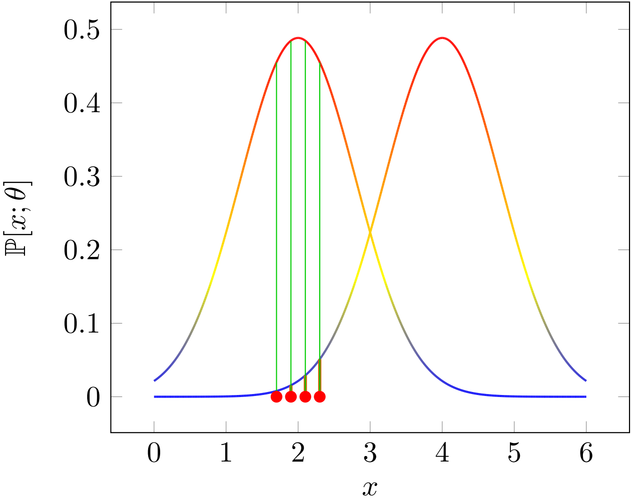
$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} x_i.$$

- Another interpretation of $\hat{\theta}$ is as the Maximum Likelihood Estimator. We first write the probability of generating the sample $\mathcal{S}$:

$$\mathbb{P}[\mathcal{S} = (x_1, x_2, \ldots, x_m)] = \prod_{i=1}^{m} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}$$

We have to find θ such that $\mathbb{P}[\mathcal{S} = (x_1, x_2, \ldots, x_m)]$ is as maximum as possible. That is,

$$\hat{\theta} = \arg\max_{\theta} \mathbb{P}[\mathcal{S}; \theta]$$



Maximum Likelihood Estimator

- We define the log likelihood of $\mathcal{S}$, given the parameter θ , as the log of the preceding expression:

$$\begin{aligned} L(\mathcal{S};\theta) &= \log\left(\mathbb{P}[\mathcal{S}=(x_1,x_2,\ldots,x_m)]\right) \\ &= \log(\theta)\sum_{i=1}^{m} x_i + \log(1-\theta)\sum_{i=1}^{m}(1-x_i) \end{aligned}$$

- The maximum likelihood estimator is the parameter that maximizes the likelihood

$$\hat{\theta} \in \arg\max_{\theta} L(\mathcal{S};\theta).$$

$$\frac{\sum_{i=1}^{m} x_i}{\theta} - \frac{\sum_{i=1}^{m}(1-x_i)}{1-\theta} = 0 \Rightarrow \hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} x_i.$$

Maximum Likelihood Estimator

- Given an IID training set $\mathcal{S} = (x_1, \ldots, x_m)$ sampled according to a density distribution $\mathcal{P}_\theta$, we define the likelihood of $\mathcal{S}$ given θ as

$$L(\mathcal{S};\theta) = \log\left(\prod_{i=1}^{m} \mathcal{P}_\theta(x_i)\right) = \sum_{i=1}^{m} \log(\mathcal{P}_\theta(x_i)).$$

- As before, the maximum likelihood estimator is a maximizer of $L(\mathcal{S};\theta)$ with respect to θ.

$$\hat{\theta} \in \arg\max_{\theta} L(\mathcal{S};\theta).$$

- As an example, consider a Gaussian random variable, for which the density function of X is parameterized by $\theta = (\mu, \sigma)$ and is defined as:

$$\mathcal{P}_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Maximum Likelihood Estimator

- We can rewrite the likelihood as:

$$L(\mathcal{S};\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - m \times \log(\sigma\sqrt{2\pi}).$$

$$\frac{d}{d\mu}L(\mathcal{S};\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu) = 0$$
$$\frac{d}{d\sigma}L(\mathcal{S};\theta) = \frac{1}{\sigma^3}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{\sigma} = 0$$

- Solving the preceding equations we obtain the maximum likelihood estimates:

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}x_i \text{ and } \hat{\sigma} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i-\hat{\mu})^2}.$$

Bayes Optimal Classifier

- The Naive Bayes classifier is a classical demonstration of how generative assumptions and parameter estimations simplify the learning process.
- Consider the problem of predicting a label $y \in \{0,1\}$ on the basis of a vector of features $\mathbf{x} = (x_1, \ldots, x_d)$, where we assume that each x_i is in $\{0,1\}$.
- The Bayes optimal classifier is

$$h_{\text{Bayes}}(\mathbf{x}) = \underset{y \in \{0,1\}}{\arg\max} \mathcal{P}[Y = y | X = \mathbf{x}].$$

- To describe the probability function $\mathcal{P}[Y = y | X = \mathbf{x}]$ we need 2^d parameters, each of which corresponds to $\mathcal{P}[Y = 1 | X = \mathbf{x}]$ for a certain value of $\mathbf{x} \in \{0,1\}^d$.
- This implies that the number of examples we need grows exponentially with the number of features.

Naive Bayes Classifier

- In the Naive Bayes approach we make the generative assumption that given the label, the features are independent of each other. That is,

$$\mathcal{P}[X = \mathbf{x}|Y = y] = \prod_{i=1}^{d} \mathcal{P}[X_i = x_i|Y = y].$$

- Now, using the Bayes rule, we have that

$$\begin{aligned} h_{\mathsf{Bayes}}(\mathbf{x}) &= \arg\max_{y\in\{0,1\}} \mathcal{P}[Y = y|X = \mathbf{x}] \\ &= \arg\max_{y\in\{0,1\}} \mathcal{P}[Y = y]\mathcal{P}[X = \mathbf{x}|Y = y] \\ &= \arg\max_{y\in\{0,1\}} \mathcal{P}[Y = y] \prod_{i=1}^{d} \mathcal{P}[X_i = \mathbf{x}_i|Y = y]. \end{aligned}$$

- That is, now the number of parameters we need to estimate is only $2d + 1$.

Linear Discriminant Analysis

- Consider the problem of predicting a label $y \in \{0, 1\}$ on the basis of a vector of features $\mathbf{x} = (x_1, \ldots, x_d)$, where we assume that each x_i is in $\{0, 1\}$.
- We assume that $\mathcal{P}[Y = 0] = \mathcal{P}[Y = 1] = \frac{1}{2}$.
- Second, we assume that the conditional probability of X given Y is a Gaussian distribution.
- The covariance matrix of the Gaussian distribution is the same for both values of the label.
- Formally, let $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathbb{R}^d$ and let $\boldsymbol{\Sigma}$ be a covariance matrix. Then, the density distribution is given by

$$\mathcal{P}[X = \mathbf{x} | Y = y] = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_y)}.$$

Linear Discriminant Analysis

- Now, using the Bayes rule we can write

$$h_{\textsf{Bayes}}(\mathbf{x}) = \underset{y\in\{0,1\}}{\arg\max} \mathcal{P}[Y=y]\mathcal{P}[X=\mathbf{x}|Y=y]$$

- This means that we will predict $h_{\textsf{Bayes}} = 1$ if and only if

$$\log\left(\frac{\mathcal{P}[Y=1]\mathcal{P}[X=\mathbf{x}|Y=1]}{\mathcal{P}[Y=0]\mathcal{P}[X=\mathbf{x}|Y=0]}\right) > 0.$$

- This ratio is often called the log-likelihood ratio. In our case, the log-likelihood ratio becomes

$$\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_0) - (\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)$$

Linear Discriminant Analysis

- We can rewrite this as $\mathbf{w}^\top \mathbf{x} + b$, where,

$$\mathbf{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1} \text{ and } b = \frac{1}{2}\left(\boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right).$$

- Under the aforementioned generative assumptions, the Bayes optimal classifier is a linear classifier.
- Additionally, one may train the classifier by estimating the parameter $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}$ from the data, using, for example, the maximum likelihood estimator.
- With those estimators at hand, the values of $\mathbf{w}$ and b can be calculated as above.