Machine Learning I: Fractal 2

Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur
http://home.iitj.ac.in/~rn/

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.

Principal Component Analysis

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

Given a dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$, reduce the dimensionality of each data-point using a linear transformation $\mathbf{W} \in \mathbb{R}^{k\times d}$, where $k < d$.

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

Algorithm 1 PCA

1: **Input:** Let $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_n\end{bmatrix}$ be a set of input points.
2: Let $\mathbf{X}\mathbf{X}^\top \mathbf{u}_i = \lambda_i \mathbf{u}_i$ be the EVD of $\mathbf{X}\mathbf{X}^\top$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.
3: $\mathbf{U} \leftarrow \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_k\end{bmatrix}$.
4: $\hat{\mathbf{x}}_i \leftarrow \mathbf{U}^\top \mathbf{x}_i, \forall i \in \{1, 2, \dots, n\}$.

Problem Formulation

$$\min_{\mathbf{H}\in\mathbb{R}^{n\times k},\mathbf{H}^\top\mathbf{H}=\mathbf{I}} \text{trace}(\mathbf{H}^\top\mathbf{L}\mathbf{H}).$$

Here, the matrix $\mathbf{L}$ is a symmetric matrix.

Rayleigh quotient

$$\mathbf{v}^\star = \arg\min_{\mathbf{v}\in\mathbb{R}^n,\mathbf{v}^\top\mathbf{v}=1} \mathbf{v}^\top\mathbf{L}\mathbf{v}$$

$$\begin{aligned} f(\mathbf{v}) &= \mathbf{v}^\top\mathbf{L}\mathbf{v} + \lambda(1-\mathbf{v}^\top\mathbf{v}) \\ \nabla f &= 2\mathbf{L}\mathbf{v} - 2\lambda\mathbf{v} \\ \mathbf{L}\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{v}^\top\mathbf{L}\mathbf{v} &= \lambda. \end{aligned}$$

Therefore, we have to minimize λ such that $\mathbf{L}\mathbf{v} = \lambda\mathbf{v}$. Hence, $\mathbf{v}^\star =$ eigenvector of the matrix $\mathbf{L}$ corresponding to the smallest eigenvalue $= \mathbf{u}_1$.

Rayleigh quotient

$$\mathbf{v}^\star = \underset{\mathbf{v}^\top\mathbf{v}=1,\mathbf{v}^\top\mathbf{u}_1=0}{\arg\min} \mathbf{v}^\top\mathbf{L}\mathbf{v}$$

Solution

$$\begin{aligned} f(\mathbf{v}) &= \mathbf{v}^\top\mathbf{L}\mathbf{v} + \lambda(1-\mathbf{v}^\top\mathbf{v}) \\ \nabla f &= 2\mathbf{L}\mathbf{v} - 2\lambda\mathbf{v} \\ \mathbf{L}\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{v}^\top\mathbf{L}\mathbf{v} &= \lambda \end{aligned}$$

Therefore, we have to minimize λ such that $\mathbf{L}\mathbf{v} = \lambda\mathbf{v}$ and $\mathbf{v}^\top\mathbf{u}_1 = 0$. $\mathbf{v}^\star$ = eigenvector of the matrix $\mathbf{L}$ corresponding to the second smallest eigenvalue $= \mathbf{u}_2$.

Rayleigh quotient

$$\mathbf{v}^\star = \underset{\mathbf{v}^\top\mathbf{v}=1,\mathbf{v}^\top\mathbf{u}_i=0,\forall i<k}{\arg\min} \mathbf{v}^\top\mathbf{L}\mathbf{v}$$

Solution

$$\begin{aligned} f(\mathbf{v}) &= \mathbf{v}^\top\mathbf{L}\mathbf{v} + \lambda(1-\mathbf{v}^\top\mathbf{v}) \\ \nabla f &= 2\mathbf{L}\mathbf{v} - 2\lambda\mathbf{v} \\ \mathbf{L}\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{v}^\top\mathbf{L}\mathbf{v} &= \lambda \end{aligned}$$

We have to minimize λ such that $\mathbf{L}\mathbf{v} = \lambda\mathbf{v}$ and $\mathbf{v}^\top\mathbf{u}_i = 0, \forall i < k$. $\mathbf{v}^\star$ = eigenvector of the matrix $\mathbf{L}$ corresponding to the k^{th} smallest eigenvalue $= \mathbf{u}_k$.

Rayleigh quotient

$$\arg\min_{\mathbf{v}_1,\ldots,\mathbf{v}_k} \sum_{i=1}^{k} \mathbf{v}_i^\top \mathbf{L}\mathbf{v}_i \text{ subject to } \mathbf{v}_i^\top \mathbf{v}_j = 1 \text{ if } i = j \text{ and } \mathbf{v}_i^\top \mathbf{v}_j = 0 \text{ if } i \neq j.$$

$$\begin{aligned}
f(\mathbf{v}_1, \ldots, \mathbf{v}_k) &= \sum_{i=1}^{k} \mathbf{v}_i^\top \mathbf{L}\mathbf{v}_i + \sum_{i=1}^{k} \lambda_i (1 - \mathbf{v}_i^\top \mathbf{v}_i) \\
\nabla_{\mathbf{v}_i} f &= 2\mathbf{L}\mathbf{v}_i - 2\lambda \mathbf{v}_i \\
\mathbf{L}\mathbf{v}_i &= \lambda_i \mathbf{v}_i \\
\mathbf{v}_i^\top \mathbf{L}\mathbf{v}_i &= \lambda_i
\end{aligned}$$

Therefore, we have to minimize $\sum_{i=1}^{k} \lambda_i$ such that $\mathbf{L}\mathbf{v}_i = \lambda \mathbf{v}_i$ and $\mathbf{v}_i^\top \mathbf{v}_j = 0$ if $i \neq j$. Hence, $\mathbf{v}_i^\star =$ eigenvector of the matrix $\mathbf{L}$ corresponding to the i^{th} smallest eigenvalue $= \mathbf{v}_i$.

Problem

$$\mathbf{H}^\star = \underset{\mathbf{H}\in\mathbb{R}^{n\times k}, \mathbf{H}^\top\mathbf{H}=\mathbf{I}}{\arg\min} \ \text{trace}(\mathbf{H}^\top\mathbf{L}\mathbf{H}).$$

Solution

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

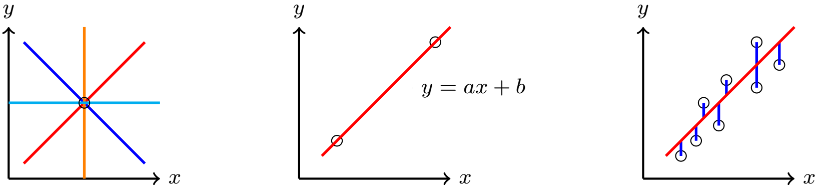
Problem

$$\mathbf{U}^\star = \underset{\mathbf{U}\in\mathbb{R}^{d\times k}, \mathbf{U}^\top\mathbf{U}=\mathbf{I}}{\arg\max} \ \text{trace}(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}).$$

Solution

Let $\mathbf{X}\mathbf{X}^\top\mathbf{u}_i = \lambda_i\mathbf{u}_i, \forall i \in \{1, 2, \ldots, n\}$ be the EVD of the matrix $\mathbf{L}$. 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 $\mathbf{U}^\star = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k\end{bmatrix}$.

Linear Regression



Consider a set of m paired points $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ represented by a linear model. Let, $\hat{y}_i = ax_i + b$ is the predicted target value. Our goal is to find the optimal line parameters a and b, such that $(\hat{y}_i - y_i)^2$ is as small as possible for all the training points. Therefore, we minimize the below error with respect to a and b.

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \Rightarrow \min_{a,b} \sum_{i=1}^{m} (ax_i + b - y_i)^2 \Rightarrow \min_{\mathbf{x}} \|\mathbf{A}\mathbf{v} - \mathbf{y}\|_2^2.$$

$$\mathbf{A} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

Solution

$$f(\mathbf{v}) = \|\mathbf{A}\mathbf{v} - \mathbf{y}\|_2^2 \Rightarrow \nabla_{\mathbf{v}} f = 2\mathbf{A}^\top \mathbf{A}\mathbf{v} - 2\mathbf{A}^\top \mathbf{y} = \mathbf{0} \Rightarrow \mathbf{v}^\star = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}.$$

Linear Regression in $\mathbb{R}^d$

$$\min_{a_1,\ldots,a_d,b} \sum_{i=1}^{m} (a_1 x_{i1} + a_2 x_{i2} \cdots + a_d x_{id} + b - y_i)^2 \Rightarrow \min_{\mathbf{x}} \|\mathbf{A}\mathbf{v} - \mathbf{y}\|_2^2.$$

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \vdots & \vdots & \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \\ b \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

Feature Selection

House #/Feature	size (m^2)	# floors	#bedrooms	# windows	Garden
1	10×10	1	4	5	yes
2	10×10	2	8	10	yes
3	10×10	3	12	15	yes
4	10×10	3	10	13	yes
5	10×10	4	10	15	yes

- Let $\mathcal{X} \subset \mathbb{R}^d$ be a dataset. That is, each point is represented as a vector of d features.
- Our goal is to learn a predictor that only relies on $k << d$ features.
- Predictors that use only a small subset of features require a smaller memory footprint and can be applied faster.
- A naive approach would be to try all subsets of k out of d features and choose the subset which leads to the best performing predictor.
- However, such an exhaustive search is usually computationally intractable.

Filters

Assess individual features, independently of other features, according to some quality measure. We can then select the k features that achieve the highest score.

Pearson's Correlation Coefficient

Consider the linear regression problem. Let $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} \end{bmatrix} \in \mathbb{R}^{m \times d}$ be a matrix containing the training points. Let $\mathbf{v} = \begin{bmatrix} x_{1j} & \cdots & x_{mj} \end{bmatrix}^\top \in \mathbb{R}^m$ be a vector denoting the j^{th} mean centered feature for all the points and let $\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix}^\top \in \mathbb{R}^m$ be the mean centered values of the targets. The occurred loss that uses only the j^{th} feature would be

$$\begin{aligned}
\min_{a,b} \sum_{i=1}^{m} (a x_{ij} + b - y_i)^2 &= \min_{a,b} \|a\mathbf{v} + b\mathbf{1} - \mathbf{y}\|_2^2 \\
f(a,b) &= \|a\mathbf{v} + b\mathbf{1} - \mathbf{y}\|_2^2 \\
\frac{\partial f}{\partial a} &= 2a\mathbf{v}^\top\mathbf{v} - 2\mathbf{y}^\top\mathbf{v} \Rightarrow a^\star = \frac{\mathbf{y}^\top\mathbf{v}}{\mathbf{v}^\top\mathbf{v}} \\
\frac{\partial f}{\partial b} &= 2b\mathbf{1}^\top\mathbf{1} \Rightarrow b^\star = 0
\end{aligned}$$

Pearson's correlation coefficient

The solution to this optimization problem is $b^\star = 0$ and $a^\star = \frac{\mathbf{v}^\top\mathbf{y}}{\mathbf{v}^\top\mathbf{v}}$. Plugging this value back into the objective we obtain the value

$$\begin{aligned}
f(a^\star, b^\star) &= \|a^\star\mathbf{v} - \mathbf{y}\|_2^2 = (a^\star\mathbf{v} - \mathbf{y})^\top(a^\star\mathbf{v} - \mathbf{y}) \\
&= (a^\star)^2\mathbf{v}^\top\mathbf{v} - 2a^\star\mathbf{y}^\top\mathbf{v} + \mathbf{y}^\top\mathbf{y} \\
&= \frac{(\mathbf{v}^\top\mathbf{y})^2}{\mathbf{v}^\top\mathbf{v}} - 2\frac{(\mathbf{v}^\top\mathbf{y})^2}{\mathbf{v}^\top\mathbf{v}} + \mathbf{y}^\top\mathbf{y} \\
&= \|\mathbf{y}\|_2^2 - \frac{(\mathbf{v}^\top\mathbf{y})^2}{\|\mathbf{v}\|_2^2} = \mathbf{y}^\top\mathbf{y} - \frac{(\mathbf{v}^\top\mathbf{y})^2}{\mathbf{v}^\top\mathbf{v}} \\
&= \|\mathbf{y}\|_2^2\left(1 - \frac{(\mathbf{v}^\top\mathbf{y})^2}{\|\mathbf{v}\|_2^2 \times \|\mathbf{y}\|_2^2}\right).
\end{aligned}$$

Ranking features according to the minimal loss is equivalent to ranking them according to the absolute value of the following score (where a higher score yields a better feature):

$$\frac{(\mathbf{v}^\top\mathbf{y})}{\|\mathbf{v}\|_2 \times \|\mathbf{y}\|_2} = \frac{\frac{1}{m}(\mathbf{v}^\top\mathbf{y})}{\sqrt{\frac{1}{m}\|\mathbf{v}\|_2^2}\sqrt{\frac{1}{m}\|\mathbf{y}\|_2^2}}$$

$$f^{\star} = \|\mathbf{y}\|_2^2 \left(1 - \frac{\frac{1}{m}(\mathbf{v}^{\top}\mathbf{y})}{\sqrt{\frac{1}{m}\|\mathbf{v}\|_2^2}\sqrt{\frac{1}{m}\|\mathbf{y}\|_2^2}}\right)$$

- The numerator is the empirical estimate of the covariance of the j-th feature and the target value, while the denominator is the squared root of the empirical estimate for the variance of the j-th feature, times the variance of the target.
- Pearson's coefficient ranges from -1 to $+1$, where if the Pearson's coefficient is either $+1$ or -1, there is a linear mapping from $\mathbf{v}$ to $\mathbf{y}$ with zero empirical risk.
- If Pearson's coefficient equals zero it means that the optimal linear function from $\mathbf{v}$ to $\mathbf{y}$ is the all-zeros function, which means that $\mathbf{v}$ alone is useless for predicting $\mathbf{y}$.
- However, this does not mean that $\mathbf{v}$ is a bad feature, as it might be the case that together with other features $\mathbf{v}$ can perfectly predict $\mathbf{y}$.