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.

Feature Selection: Filtering Techniques

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

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

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 absolute score yields a better feature):

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

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

Example

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
$x_{1,i}$	0	1	0	2	1	-1	0	-2	-1
$x_{2,i}$	0	0	1	1	2	0	-1	-1	-2
y_i	0	-3	-1	-7	-5	3	1	7	5

$$
\begin{aligned}
\|\mathbf{v}_1\|_2 &= \sqrt{1^2+2^2+1^2+(-1)^2+(-2)^2+(-1)^2} = \sqrt{12} \\
\|\mathbf{v}_2\|_2 &= \sqrt{1^2+1^2+2^2+(-1)^2+(-1)^2+(-2)^2} = \sqrt{12} \\
\|\mathbf{y}\|_2 &= \sqrt{(-3)^2+(-1)^2+(-7)^2+(-5)^2+3^2+7^2+5^2} = \sqrt{168} \\
\mathbf{v}_1^\top \mathbf{y} &= -1\times 3 - 2\times 7 - 1\times 5 - 1\times 3 - 2\times 7 - 1\times 5 = -47 \\
\mathbf{v}_2^\top \mathbf{y} &= -1\times 1 - 1\times 7 - 2\times 5 - 1\times 1 - 1\times 7 - 2\times 5 = -36 \\
\rho_1 &= \frac{\mathbf{v}_1^\top \mathbf{y}}{\|\mathbf{v}_1\|_2 \times \|\mathbf{y}\|_2} = \frac{-47}{\sqrt{168}\sqrt{12}} \\
\rho_2 &= \frac{\mathbf{v}_2^\top \mathbf{y}}{\|\mathbf{v}_2\|_2 \times \|\mathbf{y}\|_2} = \frac{-36}{\sqrt{168}\sqrt{12}} \\
\Rightarrow |\rho_1| &> |\rho_2| \\
\Rightarrow & \quad 1^{\text{st}} \text{ feature is more important.}
\end{aligned}
$$

Sequential Feature Selector

Let $\mathcal{Y} = \{y_1, y_2, \dots, y_d\}$ be a set of d feature. Our goal is to sequentially select p feature. We should have a criterion for measuring the importance of the feature. Let $\mathcal{X} \subset \mathcal{Y}$ be a subset of features. Let us define the classification accuracy score $J(\mathcal{X})$.

Sequential Forward Selection

```
1: Input: $\mathcal{Y} = \{y_1, y_2, \dots, y_d\}$ and $p$
2: Output: $\mathcal{X}_k = \{x_j \in \mathcal{Y} | j = 1, 2, \dots, k\}$
3: Initialization: $\mathcal{X}_0 = \emptyset$, $k = 0$
4: $x^+ \leftarrow \underset{x \in \mathcal{Y} \setminus \mathcal{X}_k}{\arg\max}\ J(\mathcal{X}_k \cup \{x\})$
5: $\mathcal{X}_k \leftarrow \mathcal{X}_k \cup \{x^+\}$
6: $k \leftarrow k + 1$
7: goto step 4 if $k < p$.
```

Example

- Let $\mathcal{Y} = \{y_1, y_2, y_3, y_4\}$ and $k = 2$
- If $J(y_1) = 0.3$, $J(y_2) = 0.35$, $J(y_3) = 0.45$, $J(y_1) = 0.4$ then $\mathcal{X}_1 = \{y_3\}$.
- If $J(\{y_3, y_1\}) = 0.6$, $J(\{y_3, y_2\}) = 0.7$, $J(\{y_3, y_4\}) = 0.5$ then $\mathcal{X}_2 = \{y_3, y_2\}$.

Sequential Backward Selection

```
1: Input: $\mathcal{Y} = \{y_1, y_2, \dots, y_d\}$ and $p$.
2: Output: $\mathcal{X}_k = \{x_j \in \mathcal{Y} | j = 1, 2, \dots, k\}$
3: Initialization: $\mathcal{X}_0 = \mathcal{Y}$, $k = d$.
4: $x^- \leftarrow \underset{x \in \mathcal{X}_k}{\arg\max}\ J(\mathcal{X}_k \setminus \{x\})$
5: $\mathcal{X}_k \leftarrow \mathcal{X}_k \setminus \{x^-\}$
6: $k \leftarrow k - 1$
7: goto step 4 if $k > p$.
```

Example

- Let $\mathcal{Y} = \{y_1, y_2, y_3, y_4\}$ and $k = 2$
- If $J(y_1, y_2, y_3) = 0.3$, $J(y_1, y_2, y_4) = 0.35$, $J(y_1, y_3, y_4) = 0.45$, $J(y_2, y_3, y_4) = 0.55$ then $\mathcal{X}_1 = \{y_2, y_3, y_4\}$.
- If $J(\{y_2, y_3\}) = 0.7$, $J(\{y_2, y_4\}) = 0.6$, $J(\{y_3, y_4\}) = 0.5$ then $\mathcal{X}_2 = \{y_2, y_3\}$.

Sequential Floating Selection

- Additional exclusion (inclusion) step to remove features once they were included (or excluded)
- A larger number of feature subset combinations can be sampled with this approach.

Sequential Forward Floating Selection (SFFS)

```
1:  **Input:** $\mathcal{Y} = \{y_1, y_2, \ldots, y_d\}$, required features $p$.
2:  **Output:** $\mathcal{X}_k = \{x_j \in \mathcal{Y} | j = 1, 2, \ldots, k\}$
3:  **Initialization:** $\mathcal{X}_0 = \emptyset$, $k = 0$.
4:  $x^+ \leftarrow \underset{x \in \mathcal{Y} \setminus \mathcal{X}_k}{\arg\max}\ J(\mathcal{X}_k \cup \{x\})$
5:  $\mathcal{X}_k \leftarrow \mathcal{X}_k \cup \{x^+\}$
6:  $k \leftarrow k + 1$
7:  $x^- \leftarrow \underset{x \in \mathcal{X}_k}{\arg\max}\ J(\mathcal{X}_k \setminus \{x\})$
8:  **if** $J(\mathcal{X}_k \setminus \{x\}) > J(\mathcal{X}_k)$ **then**
9:      $\mathcal{X}_k \leftarrow \mathcal{X}_k \setminus \{x^-\}$
10:     $k \leftarrow k - 1$
11: **end if**
12: goto step 4 if $k < p$.
```

Sequential Floating Selection

Sequential Backward Floating Selection (SBFS)

```
1:  **Input:** $\mathcal{Y} = \{y_1, y_2, \ldots, y_d\}$, required features $p$.
2:  **Output:** $\mathcal{X}_k = \{x_j \in \mathcal{Y} | j = 1, 2, \ldots, k\}$
3:  **Initialization:** $\mathcal{X}_0 = \mathcal{Y}$, $k = d$.
4:  $x^- \leftarrow \arg\max_{x \in \mathcal{X}_k} J(\mathcal{X}_k \backslash \{x\})$
5:  $\mathcal{X}_k \leftarrow \mathcal{X}_k \backslash \{x^-\}$
6:  $k \leftarrow k - 1$
7:  $x^+ \leftarrow \arg\max_{x \in \mathcal{Y} \backslash \mathcal{X}_k} J(\mathcal{X}_k \cup \{x\})$
8:  **if** $J(\mathcal{X}_k \cup \{x\}) > J(\mathcal{X}_k)$ **then**
9:      $\mathcal{X}_k \leftarrow \mathcal{X}_k \cup \{x^+\}$
10:     $k \leftarrow k + 1$
11: **end if**
12: goto step 4 if $k < p$.
```

Feature Transformation

We denote by $\mathbf{f} = \begin{bmatrix} f_1 & f_2 & \cdots & f_m \end{bmatrix}^\top \in \mathbb{R}^m$ the value of the feature f over the m training examples. We denote by $\bar{f} = \frac{1}{m}\sum_{i=1}^{m} f_i$ the empirical mean of the feature over all examples.
Centering: It makes the feature have zero mean, by setting $f_i \leftarrow f_i - \bar{f}$.
Unit Range: It makes the range of each feature to $[0, 1]$. Let $f_{\text{max}} = \max\{f_1, f_2, \ldots, f_m\}$ and $f_{\text{min}} = \min\{f_1, f_2, \ldots, f_m\}$. Then, we set

$$f_i \leftarrow \frac{f_i - f_{\text{min}}}{f_{\text{max}} - f_{\text{min}}}$$

Standardization: It makes all features have a zero mean and unit variance. Let $\sigma_{\mathbf{v}}^2 = \frac{1}{m}\sum\limits_{i=1}^{m}(f_i - \bar{f})^2$ be the empirical variance of the feature. Then, we set:

$$f_i \leftarrow \frac{f_i - \bar{f}}{\sigma_{\mathbf{v}}}.$$

Feature Learning

We start with some instance space, $\mathcal{X}$, and would like to learn a function, $\phi : \mathcal{X} \to \mathbb{R}^k$, which maps instances in $\mathcal{X}$ into a k-dimensional feature vectors.

Auto Encoders

We learn a pair of functions: an "encoder" function $\psi : \mathbb{R}^d \to \mathbb{R}^k$, and a "decoder" function $\phi : \mathbb{R}^k \to \mathbb{R}^d$. The goal of the learning process is to find a pair (ψ, ϕ) of functions such that the reconstruction error, defined as below, is as small as possible.

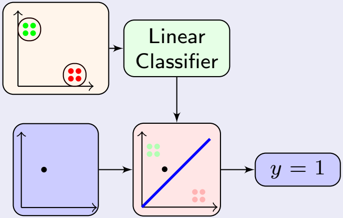
$$\sum_{i=1}^{m} \|\mathbf{x}_i - \phi(\psi(\mathbf{x}_i))\|_2^2.$$

PCA

We constrain $k < d$ and restrict ψ to a matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$ and ϕ to a matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$ and minimize the reconstruction error $\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{x}_i\|_2^2$.

Discriminative Models

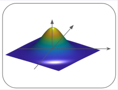
- Given a paired training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, our goal is to find a predictor h such that $h(\mathbf{x}_i)$ is equal to the true label of the input $\mathbf{x}_i$.
- We do not impose any assumptions on the underlying distribution over the data.
- Our goal is not to learn the underlying distribution but rather to learn an accurate predictor.



Generative Models

- We assumed that the underlying distribution over the data has a specific parametric form and our goal is to estimate the parameters of the model.
- The discriminative approach has the advantage of directly optimizing the quantity of interest (the prediction accuracy) instead of learning the underlying distribution.
- However, in some situations, it is reasonable to adopt the generative learning approach.

Samples from $p_{\text{data}}(\mathbf{x})$

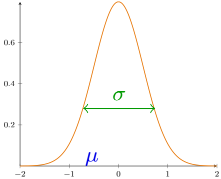


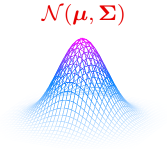
$p_\theta(\mathbf{x})$

Parameter Estimation

Given a dataset $\mathcal{S} = \{x_1, x_2, \ldots, x_m\}$ containing independent samples drawn from an unknown data distribution $p_{\text{data}}(x)$ (IID samples), learn a distribution $p_\theta(x)$ that is close as possible to the true distribution $p_{\text{data}}(x)$. Let us assume that the $p_\theta(x)$ is parametrized by the parameter θ. We have to find θ such that $p_\theta(x)$ that is close as possible to the true distribution $p_{\text{data}}(x)$.

- Parametric distributions are fully specified by a fixed number of parameters.
- E.g., Gaussian distributions are defined by the mean and covariance parameters.



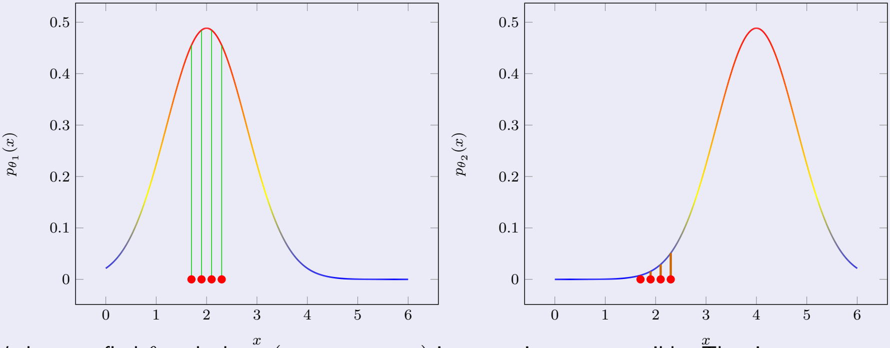


Gaussian Distribution

- $x \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow p(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}}$.

Maximum Likelihood Estimator

Define the likelihood of the data $\mathcal{S}$ with respect to $p_\theta(x)$ as: $p_\theta(x_1, x_2, \ldots, x_m)$.



We have to find θ such that $p(x_1, x_2, \ldots, x_m)$ is as maximum as possible. That is:

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta p_\theta(x_1, x_2, \ldots, x_m) = \arg\max_\theta \ell(\mathcal{S}; \theta) \\
&= \arg\max_\theta p_\theta(x_1) p_\theta(x_2) \cdots p_\theta(x_m) = \arg\max_\theta \prod_{i=1}^{m} p_\theta(x_i) \\
&= \arg\max_\theta \log\left(\prod_{i=1}^{m} p_\theta(x_i)\right) = \arg\max_\theta \sum_{i=1}^{m} \log p_\theta(x_i)
\end{aligned}$$

MLE: Example

- Assume that the samples of $\mathcal{S}$ are IID and drawn from a Gaussian distribution parameterized by $\theta = (\mu, \sigma)$.
- We can write the log-likelihood as:

$$
\begin{aligned}
\ell(\mathcal{S};\theta) &= \sum_{i=1}^{m} \log p_\theta(x_i) = \sum_{i=1}^{m} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - m \times \log(\sigma\sqrt{2\pi}) \\
\frac{\partial}{\partial\mu}\ell(\mathcal{S};\theta) &= \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu) = 0 \\
\hat{\mu} &= \frac{1}{m}\sum_{i=1}^{m} x_i \\
\frac{\partial}{\partial\sigma}\ell(\mathcal{S};\theta) &= \frac{1}{\sigma^3}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{\sigma} = 0 \\
\hat{\sigma} &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i-\hat{\mu})^2}.
\end{aligned}
$$