Machine Learning I: Fractal 2

Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur

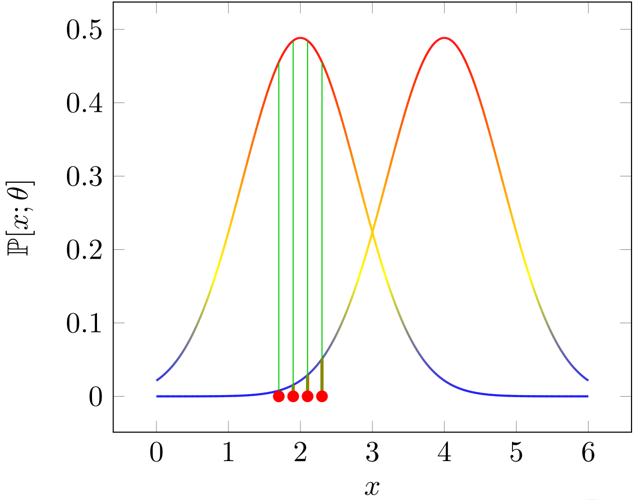
These slides are prepared from the following book:
Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

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.

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,

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



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

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

- 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

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

- This means that we will predict $h_{\text{Bayes}} = 1$ 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

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

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.

Latent Variables and the EM Algorithm

- In generative models we assume that the data is generated by sampling from a specific parametric distribution over our instance space $\mathcal{X}$.
- Sometimes, it is convenient to express this distribution using latent random variables.
- A natural example is a mixture of k Gaussian distributions.
- That is, $\mathcal{X} = \mathbb{R}^d$ and we assume that each $\mathbf{x}$ is generated as follows.
- First, we choose a random number in $1, \ldots, k$. Let Y be a random variable corresponding to this choice, and denote $P[Y = y] = c_y$.
- Second, we choose $\mathbf{x}$ on the basis of the value of Y according to a Gaussian distribution

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

Latent Variables and the EM Algorithm

- Therefore, the density of X can be written as:

$$\begin{aligned}\mathcal{P}[X = \mathbf{x}] &= \sum_{y=1}^{k} \mathcal{P}[Y = y]\mathcal{P}[X = \mathbf{x}|Y = y] \\ &= \sum_{y=1}^{k} c_y \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_y)^{\top}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_y)}\end{aligned}$$

- Note that Y is a hidden variable that we do not observe in our data. Nevertheless, we introduce Y since it helps us describe a simple parametric form of the probability of X.
- More generally, let $\boldsymbol{\theta}$ be the parameters of the joint distribution of X and Y (e.g., in the preceding example, $\boldsymbol{\theta}$ consists of c_y, $\boldsymbol{\mu}_y$, and $\mathbf{\Sigma}_y$, for all $y = 1, \ldots, k$.

Latent Variables and the EM Algorithm

- Then, the log-likelihood of an observation $\mathbf{x}$ can be written as

$$\log(\mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}]) = \log\left(\sum_{y=1}^{k} \mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}, Y = y]\right)$$

- Given an IID sample, $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, we would like to find $\boldsymbol{\theta}$ that maximizes the log-likelihood of $\mathcal{S}$.

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log\left(\prod_{i=1}^{m} \mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}_i]\right) \\ &= \sum_{i=1}^{m} \log\left(\mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}_i]\right) \\ &= \sum_{i=1}^{m} \log\left(\sum_{y=1}^{k} \mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}_i, Y = y]\right) \end{aligned}$$

Latent Variables and the EM Algorithm

- The maximum-likelihood estimator is therefore the solution of the maximization problem

$$\arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log\left(\sum_{y=1}^{k} \mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}_i, Y = y]\right)$$

- In many situations, the summation inside the log makes the preceding optimization problem computationally hard.

Latent Variables and the EM Algorithm

- EM is designed for those cases in which, had we known the values of the latent variables Y.
- Then the maximum likelihood optimization problem would have been tractable.
- More precisely, define the following function over $m \times k$ matrices and the set of parameters $\boldsymbol{\theta}$:

$$F(\mathbf{Q}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{Q}_{i,y} \log \left(\mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}_i, Y = y] \right)$$

- If each row of Q defines a probability over the i^{th} latent variable given $X = \mathbf{x}_i$.

Latent Variables and the EM Algorithm

- On one hand, had we known $\mathbf{Q}$, then by our assumption, the optimization problem of finding the best $\boldsymbol{\theta}$ is tractable.
- On the other hand, had we known the parameters $\boldsymbol{\theta}$ we could have set $\mathbf{Q}_{i,y}$ to be the probability of $Y = y$ given that $X = \mathbf{x}_i$
- The EM algorithm therefore alternates between finding $\boldsymbol{\theta}$ given $\mathbf{Q}$ and finding $\mathbf{Q}$ given $\boldsymbol{\theta}$. Formally, EM finds a sequence of solutions $(\mathbf{Q}^{(1)}, \boldsymbol{\theta}^{(1)})$, $(\mathbf{Q}^{(2)}, \boldsymbol{\theta}^{(2)})$, ..., where at iteration t, we construct $(\mathbf{Q}^{(t+1)}, \boldsymbol{\theta}^{(t+1)})$ by performing two steps.

Expectation Step

Set

$$\mathbf{Q}^{(t+1)} = \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y = y | X = \mathbf{x}_i]$$

This step is called the Expectation step, because it yields a new probability over the latent variables, which defines a new expected log-likelihood function over $\boldsymbol{\theta}$.

Minimization Step

Set

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} F(\mathbf{Q}^{t+1,\boldsymbol{\theta}})$$

This step is called the Expectation step, because it yields a new probability over the latent variables, which defines a new expected log-likelihood function over $\boldsymbol{\theta}$.