vector of features $\mathbf{x} = (x_1, \ldots, x_d)$. But now the generative assumption is as follows. First, we assume that $\mathcal{P}[Y = 1] = \mathcal{P}[Y = 0] = 1/2$. Second, we assume that the conditional probability of X given Y is a Gaussian distribution. Finally, 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 Σ be a covariance matrix. Then, the density distribution is given by

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

As we have shown in the previous section, using Bayes' rule we can write

$$h_{\text{Bayes}}(\mathbf{x}) = \operatorname*{argmax}_{y \in \{0,1\}} \mathcal{P}[Y = y]\mathcal{P}[X = \mathbf{x}|Y = y].$$

This means that we will predict $h_{\text{Bayes}}(\mathbf{x}) = 1$ iff

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

This ratio is often called the *log-likelihood ratio*.

In our case, the log-likelihood ratio becomes

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

We can rewrite this as $\langle \mathbf{w}, \mathbf{x} \rangle + b$ where

$$\mathbf{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \Sigma^{-1} \quad \text{and} \quad b = \tfrac{1}{2}\left(\boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1\right). \tag{24.8}$$

As a result of the preceding derivation we obtain that under the aforementioned generative assumptions, the Bayes optimal classifier is a linear classifier. Additionally, one may train the classifier by estimating the parameter $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ and Σ 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 in Equation (24.8).

24.4 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 $\mathcal{P}[Y = y] = c_y$. Second, we choose $\mathbf{x}$ on the basis of the value of Y according to a Gaussian distribution

$$\mathcal{P}[X = \mathbf{x}|Y = y] = \frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T \Sigma_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)\right). \tag{24.9}$$

Therefore, the density of X can be written as:

$$\begin{aligned}\mathcal{P}[X = \mathbf{x}] &= \sum_{y=1}^{k} \mathcal{P}[Y = y]\mathcal{P}[X = \mathbf{x}|Y = y] \\ &= \sum_{y=1}^{k} c_y \frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T \Sigma_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)\right).\end{aligned}$$

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 Σ_y, for all $y = 1, \ldots, k$). Then, the log-likelihood of an observation $\mathbf{x}$ can be written as

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

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

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

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

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

In many situations, the summation inside the log makes the preceding optimization problem computationally hard. The *Expectation-Maximization* (EM) algorithm, due to Dempster, Laird, and Rubin, is an iterative procedure for searching a (local) maximum of $L(\boldsymbol{\theta})$. While EM is not guaranteed to find the global maximum, it often works reasonably well in practice.

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 $\boldsymbol{\theta}$:

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

If each row of Q defines a probability over the ith latent variable given $X = \mathbf{x}_i$, then we can interpret $F(Q, \boldsymbol{\theta})$ as the expected log-likelihood of a training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, where the expectation is with respect to the choice of each y_i on the basis of the ith row of Q. In the definition of F, the summation is outside the log, and we assume that this makes the optimization problem with respect to $\boldsymbol{\theta}$ tractable:

ASSUMPTION 24.1 For any matrix $Q \in [0,1]^{m,k}$, such that each row of Q sums to 1, the optimization problem

$$\operatorname*{argmax}_{\boldsymbol{\theta}} F(Q, \boldsymbol{\theta})$$

is tractable.

The intuitive idea of EM is that we have a "chicken and egg" problem. On one hand, had we known 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 $Q_{i,y}$ to be the probability of $Y = y$ given that $X = \mathbf{x}_i$. The EM algorithm therefore alternates between finding $\boldsymbol{\theta}$ given Q and finding Q given $\boldsymbol{\theta}$. Formally, EM finds a sequence of solutions $(Q^{(1)}, \boldsymbol{\theta}^{(1)}), (Q^{(2)}, \boldsymbol{\theta}^{(2)}), \ldots$ where at iteration t, we construct $(Q^{(t+1)}, \boldsymbol{\theta}^{(t+1)})$ by performing two steps.

- **E**xpectation Step: Set

$$Q_{i,y}^{(t+1)} = \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y = y | X = \mathbf{x}_i]. \tag{24.10}$$

  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 $\boldsymbol{\theta}$.
- **M**aximization Step: Set $\boldsymbol{\theta}^{(t+1)}$ to be the maximizer of the expected log-likelihood, where the expectation is according to $Q^{(t+1)}$:

$$\boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} F(Q^{(t+1)}, \boldsymbol{\theta}). \tag{24.11}$$

  By our assumption, it is possible to solve this optimization problem efficiently.

The initial values of $\boldsymbol{\theta}^{(1)}$ and $Q^{(1)}$ are usually chosen at random and the procedure terminates after the improvement in the likelihood value stops being significant.

24.4.1 EM as an Alternate Maximization Algorithm

To analyze the EM algorithm, we first view it as an alternate maximization algorithm. Define the following objective function

$$G(Q, \boldsymbol{\theta}) = F(Q, \boldsymbol{\theta}) - \sum_{i=1}^{m} \sum_{y=1}^{k} Q_{i,y} \log(Q_{i,y}).$$

The second term is the sum of the *entropies* of the rows of Q. Let

$$\mathbb{Q} = \left\{ Q \in [0,1]^{m,k} : \forall i,\ \sum_{y=1}^{k} Q_{i,y} = 1 \right\}$$

be the set of matrices whose rows define probabilities over $[k]$. The following lemma shows that EM performs alternate maximization iterations for maximizing G.

LEMMA 24.2 *The EM procedure can be rewritten as:*

$$\begin{aligned} Q^{(t+1)} &= \operatorname*{argmax}_{Q \in \mathbb{Q}} G(Q, \boldsymbol{\theta}^{(t)}) \\ \boldsymbol{\theta}^{(t+1)} &= \operatorname*{argmax}_{\boldsymbol{\theta}} G(Q^{(t+1)}, \boldsymbol{\theta}) \ . \end{aligned}$$

Furthermore, $G(Q^{(t+1)}, \boldsymbol{\theta}^{(t)}) = L(\boldsymbol{\theta}^{(t)})$.

Proof Given $Q^{(t+1)}$ we clearly have that

$$\operatorname*{argmax}_{\boldsymbol{\theta}} G(Q^{(t+1)}, \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} F(Q^{(t+1)}, \boldsymbol{\theta}).$$

Therefore, we only need to show that for any $\boldsymbol{\theta}$, the solution of $\operatorname{argmax}_{Q \in \mathbb{Q}} G(Q, \boldsymbol{\theta})$ is to set $Q_{i,y} = \mathcal{P}_{\boldsymbol{\theta}}[Y = y | X = \mathbf{x}_i]$. Indeed, by Jensen's inequality, for any $Q \in \mathbb{Q}$ we have that

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

while for $Q_{i,y} = \mathcal{P}_{\boldsymbol{\theta}}[Y = y|X = \mathbf{x}_i]$ we have

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

This shows that setting $Q_{i,y} = \mathcal{P}_{\boldsymbol{\theta}}[Y = y|X = \mathbf{x}_i]$ maximizes $G(Q, \boldsymbol{\theta})$ over $Q \in \mathbb{Q}$ and shows that $G(Q^{(t+1)}, \boldsymbol{\theta}^{(t)}) = L(\boldsymbol{\theta}^{(t)})$. □

The preceding lemma immediately implies:

THEOREM 24.3 *The EM procedure never decreases the log-likelihood; namely, for all t,*

$$L(\boldsymbol{\theta}^{(t+1)}) \geq L(\boldsymbol{\theta}^{(t)}).$$

Proof By the lemma we have

$$L(\boldsymbol{\theta}^{(t+1)}) = G(Q^{(t+2)}, \boldsymbol{\theta}^{(t+1)}) \geq G(Q^{(t+1)}, \boldsymbol{\theta}^{(t)}) = L(\boldsymbol{\theta}^{(t)}).$$

□

24.4.2 EM for Mixture of Gaussians (Soft k-Means)

Consider the case of a mixture of k Gaussians in which $\boldsymbol{\theta}$ is a triplet $(\mathbf{c}, \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}, \{\Sigma_1, \ldots, \Sigma_k\})$ where $\mathcal{P}_{\boldsymbol{\theta}}[Y = y] = c_y$ and $\mathcal{P}_{\boldsymbol{\theta}}[X = \mathbf{x}|Y = y]$ is as given in Equation (24.9). For simplicity, we assume that $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = I$, where I is the identity matrix. Specifying the EM algorithm for this case we obtain the following:

- **E**xpectation step: For each $i \in [m]$ and $y \in [k]$ we have that

$$\begin{aligned} \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y = y|X = \mathbf{x}_i] &= \frac{1}{Z_i} \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y = y] \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[X = \mathbf{x}_i|Y = y] \\ &= \frac{1}{Z_i} c_y^{(t)} \exp\left(-\frac{1}{2} \|\mathbf{x}_i - \boldsymbol{\mu}_y^{(t)}\|^2 \right), \end{aligned} \tag{24.12}$$

  where Z_i is a normalization factor which ensures that $\sum_y \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y = y|X = \mathbf{x}_i]$ sums to 1.
- **M**aximization step: We need to set $\boldsymbol{\theta}^{t+1}$ to be a maximizer of Equation (24.11),

which in our case amounts to maximizing the following expression w.r.t. $\mathbf{c}$ and $\boldsymbol{\mu}$:

$$\sum_{i=1}^{m}\sum_{y=1}^{k} \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y=y|X=\mathbf{x}_i]\left(\log(c_y) - \frac{1}{2}\|\mathbf{x}_i - \boldsymbol{\mu}_y\|^2\right). \tag{24.13}$$

Comparing the derivative of Equation (24.13) w.r.t. $\boldsymbol{\mu}_y$ to zero and rearranging terms we obtain:

$$\boldsymbol{\mu}_y = \frac{\sum_{i=1}^{m} \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y=y|X=\mathbf{x}_i]\,\mathbf{x}_i}{\sum_{i=1}^{m} \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y=y|X=\mathbf{x}_i]}.$$

That is, $\boldsymbol{\mu}_y$ is a weighted average of the $\mathbf{x}_i$ where the weights are according to the probabilities calculated in the E step. To find the optimal $\mathbf{c}$ we need to be more careful since we must ensure that $\mathbf{c}$ is a probability vector. In Exercise 3 we show that the solution is:

$$c_y \;=\; \frac{\sum_{i=1}^{m} \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y=y|X=\mathbf{x}_i]}{\sum_{y'=1}^{k}\sum_{i=1}^{m} \mathcal{P}_{\boldsymbol{\theta}^{(t)}}[Y=y'|X=\mathbf{x}_i]}. \tag{24.14}$$

It is interesting to compare the preceding algorithm to the k-means algorithm described in Chapter 22. In the k-means algorithm, we first assign each example to a cluster according to the distance $\|\mathbf{x}_i - \boldsymbol{\mu}_y\|$. Then, we update each center $\boldsymbol{\mu}_y$ according to the average of the examples assigned to this cluster. In the EM approach, however, we determine the probability that each example belongs to each cluster. Then, we update the centers on the basis of a weighted sum over the entire sample. For this reason, the EM approach for k-means is sometimes called "soft k-means."

24.5 Bayesian Reasoning

The maximum likelihood estimator follows a frequentist approach. This means that we refer to the parameter θ as a fixed parameter and the only problem is that we do not know its value. A different approach to parameter estimation is called Bayesian reasoning. In the Bayesian approach, our uncertainty about θ is also modeled using probability theory. That is, we think of θ as a random variable as well and refer to the distribution $\mathcal{P}[\theta]$ as a *prior distribution*. As its name indicates, the prior distribution should be defined by the learner prior to observing the data.

As an example, let us consider again the drug company which developed a new drug. On the basis of past experience, the statisticians at the drug company believe that whenever a drug has reached the level of clinic experiments on people, it is likely to be effective. They model this prior belief by defining a density distribution on θ such that

$$\mathcal{P}[\theta] = \begin{cases} 0.8 & \text{if } \theta > 0.5 \\ 0.2 & \text{if } \theta \leq 0.5 \end{cases} \tag{24.15}$$