Chapter *3* Generative Models

Siheng Zhang
zhangsiheng@cvte.com

November 19, 2020

This part corresponds to **Chapter 24, 31 in UML, Chapter 1, 2 in PRML**, and mainly answers the following questions:

- How to bring Bayes Optimal classifier into application? (Feature independent assumption)
- To estimate the class conditional probability distribution for Bayes classifier, we study both the parametric (*includes a family of basic probability distributions*) and non-parametric methods.
- A glance for generative and discriminant models. Naive Bayes, GMM, and etc, belong to the former, which requires estimation of underlying distribution. This is more general and hence difficult. Discriminant models try to avoid it by optimization.
- Last but not the least, there is a connection between generative and discriminant models. At last of this chapter, we point out how to derive a linear discriminant from Bayes classifier. As we will see in the next chapter, discriminant with penalization also has a intrinsic connection with generative models with some prior distribution.

Contents

1 Naive Bayes

Recall that the Bayes optimal classifier (*in Chapter 1, Ex6*) is:

$$h_{\text{Bayes}}(\boldsymbol{x}) = \arg \max_{y \in \{0,1\}} p(Y = y | X = \boldsymbol{x})$$

To describe the posterior probability function we need 2^d parameters, this implies that the number of examples we need grows exponentially with the number of features. To avoid this problem, we assume that given the label, the features are independent of each other, i.e.,

$$p(X = \boldsymbol{x} | Y = y) = \prod_{i=1}^{d} p(X_i = x_i | Y = y)$$

Together with Bayes' rule, the Bayes optimal classifier can be simplified as:

$$h_{\text{Bayes}}(\boldsymbol{x}) = \arg \max_{y \in \{0,1\}} p(Y = y) \prod_{i=1}^{d} p(X_i = x_i | Y = y) \tag{1}$$

Now the number of parameters we need to estimate is only $2d + 1$. When we also estimate the parameters using the maximum likelihood principle (see below), the resulting classifier is called the *Naive Bayes* classifier.

2 Parametric density estimation

To apply the Bayesian decision principle, assume that the form of distribution is known, the problem is to estimate the parameters. Specifically, given an i.i.d. training set $S = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$ sampled according to a density distribution, the likelihood of S given θ is $L(S; \theta) = \prod_{i=1}^{m} p(\boldsymbol{x}_i; \theta)$. Usually, we turn to optimize its logarithm,

$$\log L(S; \theta) = \sum_{i=1}^{m} \log p(\boldsymbol{x}_i; \theta) \tag{2}$$

1 Bernoulli distribution, $\theta = \mu$

Bernoulli distribution describes the probability of a binary variable x. The probability of $x = 1$ is denoted by parameter μ, and of $x = 0$ is $1 - \mu$, so, $p(x; \theta) = \mu^x (1 - \mu)^{(1-x)}$. The log likelihood function is given by

$$\log L(S; \theta) = \sum_{i=1}^{m} \log p(x_i; \theta) = \sum_{i=1}^{m} x_i \log \mu + (1 - x_i) \log(1 - \mu)$$

Set the derivative of the log likelihood with respect to μ to zero,

$$\frac{\partial \log L(S; \theta)}{\partial \mu} = \sum_{i=1}^{m} \frac{x_i}{\mu} - \frac{1 - x_i}{1 - \mu} = \sum_{i=1}^{m} \frac{x_i - \mu}{\mu(1 - \mu)} = 0 \Longrightarrow \mu_{\text{ML}} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

2 Multinomial distribution, $\theta = \boldsymbol{\mu}$

Multinomial distribution extends the binary variable to one of d possible value. The random variable can be represented by a d-dimensional vector $\boldsymbol{x}$, in which only one element equals 1 and others equal 0. Denote the probability of $x_j = 1$ by μ_j, then the distribution is given by

$$p(\boldsymbol{x} | \boldsymbol{\mu}) = \prod_{j=1}^{d} \mu_j^{x_j} \quad s.t. \quad \sum_{j=1}^{d} \mu_j = 1, \ \forall j, \quad \mu_j \geq 0$$

The corresponding log likelihood function is given by

$$\log L(S; \theta) = \sum_{i=1}^{m} \log p(\boldsymbol{x}_i; \theta) = \sum_{i=1}^{m} \sum_{j=1}^{d} x_{ij} \log \mu_j$$

Using Lagrange multiplier λ, it is equivalent to maximize $L' = \log L(S;\theta) + \lambda\left(\sum_{j=1}^{d}\mu_j - 1\right)$. Set the derivative with regard to μ_j to be zero

$$\frac{\partial L'}{\partial \mu_j} = \sum_{i=1}^{m}\frac{x_{ij}}{\mu_j} + \lambda = 0 \Longrightarrow \mu_{j,\mathrm{ML}} = -\sum_{i=1}^{m} x_{ij}/\lambda$$

Besides, $\sum_{j=1}^{d}\mu_j = -m/\lambda = 1$, thereby leading to $\lambda = -m$, and $\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$.

3 Gaussian distribution, $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The Gaussian distribution is $p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$.

The log likelihood function is given by

$$\log L(S;\theta) = \sum_{i=1}^{m}\log p(\boldsymbol{x}_i;\theta) = \frac{-md}{2}\log(2\pi) - \frac{m}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

Set the derivative of the log likelihood with respect to $\boldsymbol{\mu}$ to be zero leading to $\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$.

remark1: Deriving $\boldsymbol{\Sigma}$ requires the use of the following linear algebra and calculus properties:

- The trace is invariant under cyclic permutation of matrix products: $tr[\boldsymbol{ABC}] = tr[\boldsymbol{CAB}] = tr[\boldsymbol{BCA}]$;
- Since $\boldsymbol{x}^\top\boldsymbol{Ax}$ is a scalar, its trace is itself, and hence $\boldsymbol{x}^\top\boldsymbol{Ax} = tr[\boldsymbol{x}^\top\boldsymbol{Ax}] = tr[\boldsymbol{xx}^\top\boldsymbol{A}]$;
- $\partial tr[\boldsymbol{AB}]/\partial\boldsymbol{A} = \boldsymbol{B}^\top$; $\partial\log|\boldsymbol{A}|/\partial\boldsymbol{A} = (\boldsymbol{A}^{-1})^\top$; $\partial tr(\boldsymbol{AX}^{-1}\boldsymbol{B})/\partial\boldsymbol{X} = -(\boldsymbol{X}^{-1}\boldsymbol{BAX}^{-1})^\top$

The derivative of the log likelihood with respect to $\boldsymbol{\Sigma}$ is given by

$$\frac{\partial\log L(S;\theta)}{\partial\boldsymbol{\Sigma}} = -\frac{m}{2}(\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2}\sum_{i=1}^{m}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}$$

Here we does not give a formal proof that $\boldsymbol{\Sigma}$ is symmetric but directly using this conclusion, and setting the derivative to zero leads to $\boldsymbol{\Sigma}_{\mathrm{ML}} = \sum_{i=1}^{m}(\boldsymbol{x}_i-\boldsymbol{\mu}_{\mathrm{ML}})(\boldsymbol{x}_i-\boldsymbol{\mu}_{\mathrm{ML}})^\top/m$.

4 Exponential family The exponential family is defined to be the set of distributions of the form

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\{\boldsymbol{\eta}^\top\boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\eta})\} \tag{3}$$

remark1: Bernoulli distribution is a member in this family,

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp\{x\log\mu + (1-x)\log(1-\mu)\} = \exp\left\{\log\left(\frac{\mu}{1-\mu}\right)x + \log(1-\mu)\right\}$$

Compare with the general form shows that $h(x) = 1, u(x) = x, \eta = \log\frac{\mu}{1-\mu}$, and $A(\eta) = \log(1+\exp(\eta))$.

remark2: Multinomial distribution is a member in this family. Recall that multinomial distribution indeed has $d-1$ parameters since $\sum_{j=1}^{d}\mu_d = 1$, we have

$$\begin{aligned}p(\boldsymbol{x}|\boldsymbol{\mu}) &= \prod_{j=1}^{d}\mu_j^{x_j} = \exp\left\{\sum_{j=1}^{d}x_j\log\mu_j\right\} = \exp\left\{\sum_{j=1}^{d-1}x_j\log\mu_j + \left(1-\sum_{j=1}^{d-1}x_j\right)\log\left(1-\sum_{j=1}^{d-1}\mu_j\right)\right\}\\ &= \exp\left\{\sum_{j=1}^{d-1}x_j\log\left(\frac{\mu_j}{1-\sum_{k=1}^{d-1}\mu_k}\right) + \log\left(1-\sum_{j=1}^{d-1}\mu_j\right)\right\}\end{aligned}$$

Define $\eta_j = \log\frac{\mu_j}{1-\sum_{k=1}^{d}\mu_k}$, then $\mu_j = \frac{\exp\eta_j}{1+\sum_{k=1}^{d}\exp\eta_k}$, and $1-\sum_{j=1}^{d-1}\mu_j = 1 - \frac{\sum_{j=1}^{d-1}\exp\eta_j}{1+\sum_{k=1}^{d}\exp\eta_k} = \frac{\exp\eta_d}{1+\sum_{k=1}^{d}\exp\eta_k}$. Compare with the general form shows that $h(\boldsymbol{x}) = 1, u(\boldsymbol{x}) = \boldsymbol{x}, A(\boldsymbol{\eta}) = \log(1+\sum_{k=1}^{d}\exp\eta_k) - \eta_d$.

remark3: Gaussian distribution is a member in this family.

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)$$

Since $\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = tr[\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}] = tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{xx}^\top]$. Compare with the general form shows that $h(\boldsymbol{x}) = (2\pi)^{-d/2}, u(\boldsymbol{x}) = (1, \boldsymbol{x}, \boldsymbol{xx}^\top)^\top, \boldsymbol{\eta} = (-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\log|\boldsymbol{\Sigma}|, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1})^\top$.

Now consider the problem of estimating the parameter vector $\boldsymbol{\mu}$ in the general exponential family distribution. The log likelihood function is given by

$$\sum_{i=1}^{m} \log h(\boldsymbol{x}_i) + \boldsymbol{\eta}^\top \sum_{i=1}^{m} u(\boldsymbol{x}_i) - \sum_{i=1}^{m} A(\boldsymbol{\eta})$$

Take derivative with regard to $\boldsymbol{\eta}$ leads to $\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{m} u(\boldsymbol{x}_i)/m$, which can be solved to obtain $\boldsymbol{\eta}_{\mathrm{ML}}$.

Note that $\int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\eta})\} = 1$. Take derivatives of both sides with regard to $\boldsymbol{\eta}$, we have,

$$\int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\eta})\} \left(\boldsymbol{u}(\boldsymbol{x}) - \frac{\partial A(\boldsymbol{\eta})}{\partial \eta} \right) = 0$$

which leads to

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbb{E}[u(\boldsymbol{x})] \tag{4}$$

Therefore, $\sum_i u(\boldsymbol{x}_i)$ is called the sufficient statistic. Also note that the covariance of $u(\boldsymbol{x})$ can be expressed in terms of the second derivatives $A(\boldsymbol{\eta})$, and similarly for higher order moments. Thus, provided we can normalize a distribution from the exponential family, we can always find its moments by simple differentiation.

3 Non-Parametric density estimation

An important limitation of parametric approach is that the chosen density might be a poor model of the distribution that generates the data, which can result in poor predictive performance. This section considers some non-parametric approaches to density estimation that make few assumptions about the form of the distribution.

4 From MLE to Bayesian reasoning

Intuitively, MLE can give severely over-fitted results for small data sets. Formally, given a parameter $\boldsymbol{\theta}$ and an observation $\boldsymbol{x}$, define the empirical loss of $\boldsymbol{\theta}$ on $\boldsymbol{x}$ as the negative logarithm of its probability

$$l(\boldsymbol{\theta}, \boldsymbol{x}) = -\log \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x})$$

Hence, MLE is equivalent to ERM, *i.e.*,

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} -\log \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

However, the true risk of $\boldsymbol{\theta}$ according to the underlying distribution $\mathcal{P}$ is

$$\mathbb{E}[l(\boldsymbol{\theta}, \boldsymbol{x})] = -\sum_{\boldsymbol{x}} \mathcal{P}(\boldsymbol{x}) \log \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \mathcal{P}(\boldsymbol{x}) \log \left(\frac{\mathcal{P}(\boldsymbol{x})}{\mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x})} \right) + \sum_{\boldsymbol{x}} \mathcal{P}(\boldsymbol{x}) \log \frac{1}{\mathcal{P}(\boldsymbol{x})} \geq \sum_{\boldsymbol{x}} \mathcal{P}(\boldsymbol{x}) \log \frac{1}{\mathcal{P}(\boldsymbol{x})}$$

in which the equality holds *iff.* $\mathcal{P} = \mathcal{P}_{\boldsymbol{\theta}}$. In some situations, it is easy to prove that MLE guarantees low true risk. For example, consider the problem of estimating the mean of a Gaussian variable of known variance,

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} [l(\boldsymbol{\mu}_{\mathrm{ML}}, \boldsymbol{x}) - l(\boldsymbol{\mu}, \boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \log \left(\frac{\mathcal{P}_{\boldsymbol{\mu}}(\boldsymbol{x})}{\mathcal{P}_{\boldsymbol{\mu}_{\mathrm{ML}}}(\boldsymbol{x})} \right) = \frac{1}{2} (\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})$$

from which we can know that the difference of the true risk with the minimal loss is bounded.

Also, we want to know the worst case that MLE may achieve. Consider a Bernoulli random variable with parameter μ, assume that it is nonzero but very small. Then, the probability that no element of a sample of size m will be 1 is $(1-\mu)^m \geq e^{-2m\mu}$. And in that case, $\mu_{\mathrm{ML}} = 0$. But the true risk is $\mathbb{E}[l(\boldsymbol{\mu}_{\mathrm{ML}}, x)] = \mu l(\boldsymbol{\mu}_{\mathrm{ML}}, 1) + (1-\mu) l(\boldsymbol{\mu}_{\mathrm{ML}}, 0) = \theta \log(1/\mu_{\mathrm{ML}}) = \infty$.

To address this problem, we develop a Bayesian treatment, which introduce a prior distribution $p(\boldsymbol{\mu})$. To determine the prior distribution, we expect that the posterior distribution will have the same functional form as the prior. This is called **conjugacy**, and the prior is called **conjugate prior**.

1 Beta distribution for Bernoulli distribution

Recall that the likelihood of Bernoulli distribution is proportional to $\mu^x(1-\mu)^{1-x}$, we choose a prior to be proportional to powers of μ and $1-\mu$, then the posterior distribution, which is proportional to the product of the prior and the likelihood function, will have the same functional form as the prior.

The Beta distribution

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{5}$$

meets the requirement. Note that the gamma functions $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ are used to ensure the Beta distribution is normalized, so that $\int_0^1 \text{Beta}(\mu; a, b)\mathrm{d}\mu = 1$.

Given the observed sequence S,

$$p(\mu|S) \propto p(S|\mu)\text{Beta}(\mu|a,b) = \left(\prod_{i=1}^m \mu_i^x(1-\mu)^{1-x_i}\right)\mu^{a-1}(1-\mu)^{b-1} = \mu^{a+\sum_{i=1}^m x_i-1}(1-\mu)^{m-\sum_{i=1}^m x_i+b-1}$$

To ensure that it is normalized, the posterior must be $\text{Beta}(a+\sum_{i=1}^m x_i, b+m-\sum_{i=1}^m x_i)$.

Using the mean of the Beta distribution $\mathbb{E}(\mu) = \frac{a}{a+b}$, the estimated probability of a new event $x_i = 1$ is given by the mean of posterior, which

$$p(x=1|S) = \int_0^1 p(x=1|\mu)p(\mu|S)\mathrm{d}\mu = \int_0^1 \mu p(\mu|S)\mathrm{d}\mu = \mathbb{E}(\mu|S) = \frac{a+\sum_{i=1}^m x_i}{b+m}$$

Note that as the training sequence S become infinitely large, $m \to \infty$, the result convergences to $\frac{\sum_{i=1}^m x_i}{m}$, which is the same as MLE.

2 Dirichlet distribution for multinomial distribution

By inspection of the form of the multinomial distribution, the conjugate prior is given by $p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{j=1}^d \mu_j^{\alpha_j-1}$, where $0 \le \mu_k \le 1$. Its normalized form is (in which $\alpha_0 = \sum_{j=1}^d \alpha_j$):

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_d)}\prod_{j=1}^d \mu_j^{\alpha_j-1}$$

Given the observed sequence S,

$$p(\boldsymbol{\mu}|S) \propto p(S|\boldsymbol{\mu})\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \prod_{i=1}^m\prod_{j=1}^d \mu_j^{x_{ij}}\prod_{j=1}^d \mu_j^{\alpha_j-1} = \prod_{j=1}^d \mu_j^{\sum_{i=1}^m x_{ij}}\prod_{j=1}^d \mu_j^{\alpha_j-1} = \prod_{j=1}^d \mu_j^{m_j+\alpha_j-1}$$

in which we denote $m_j = \sum_{i=1}^m x_{ij}$. The normalized form of the posterior is then given by $\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m})$.

3 Gaussian distribution

There are two parameters to be estimated in Gaussian distribution, the mean vector and the covariance matrix. So there are three cases

a Known covariance, unknown mean. The conjugate prior is another Gaussian distribution $p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. The posterior is given by

$$\begin{aligned}
p(\boldsymbol{\mu}|S) &\propto p(S|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_\mathbf{0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right)\prod_{i=1}^m \exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right) \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^m(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right) \\
&= \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^\top(m\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^m \boldsymbol{x}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \sum_{i=1}^m \boldsymbol{x}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\right)\right]
\end{aligned}$$

Its normalized form is $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, in which $\boldsymbol{\Sigma}_1^{-1} = m\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^m \boldsymbol{x}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$.

b Known mean, unknown covariance. **For 1d case**, denote $\lambda = 1/\sigma^2$, the corresponding conjugate prior should therefore be proportional to the product of a power of λ and the exponential of a linear function of λ. This corresponds to the gamma distribution which is defined by

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

The posterior is given by

$$p(\lambda|S) \propto \lambda^{a_0-1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{m}^{i=1}(x_i-\mu)^2\right\}$$

For multi-variate case, the corresponding prior is Wishart distribution,

$$\mathrm{Wishart}(\boldsymbol{\Sigma}|\boldsymbol{W},v) = B|\boldsymbol{\Sigma}|^{(v-d-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{W}^{-1}\boldsymbol{\Sigma})\right)$$

c Unknown mean and covariance. The corresponding prior is Gaussian-Gamma distribution or Gaussian-Wishart distribution. We do not expand them here.

4 Exponential distribution

5 EM algorithm: MLE for partial observed data

Until now, a training sequence is $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$, in which y_i is the latent factor that depends whether $\mathrm{x_i}$ is sampled from. However, if the latent factors are not observed, the likelihood of the sequence $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ is:

$$L(S;\theta) = \prod_{i=1}^{m}\sum_{j=1}^{k} p_\theta(\boldsymbol{x}_i, y_j) = \prod_{i=1}^{m}\sum_{j=1}^{k} p_\theta(\boldsymbol{x}_i|y_j)p_\theta(y_j)$$

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

$$\log L(S;\theta) = \sum_{i=1}^{m}\log\sum_{j=1}^{k} p_\theta(\boldsymbol{x}_i|y_j)p_\theta(y_j) \tag{6}$$

In the E-step, we use the current parameter values θ^{old} to find the posterior distribution of the latent variables given by $p(\boldsymbol{Y}|\boldsymbol{X},\theta^{\mathrm{old}})$. We then use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value θ. This expectation, denoted , is given by

5.1 EM for GMM

GMM (Gaussian mixture models) is a typical example, with parameters comprising the means and covariances of the components and the mixing coefficients. Its log-likelihood function (plus a Lagrange multiplier) is given by

$$\sum_{i=1}^{m}\log\sum_{j=1}^{k}\pi_j\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) + \lambda\left(\sum_{j=1}^{k}\pi_j - 1\right)$$

Take derivatives with regard to $\boldsymbol{\mu}_k$ and set it to zero

$$\sum_{i=1}^{m}\underbrace{\frac{\pi_j\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{\sum_l \pi_l\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}}_{z_{ij}}\boldsymbol{\Sigma}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_j) \Longrightarrow \boldsymbol{\mu}_j = \frac{\sum_{i=1}^{m} z_{ij}\boldsymbol{x}_i}{\sum_{i=1}^{m} z_{ij}} \tag{7}$$

in which $z_{ij} = p(y_j = 1|\boldsymbol{x}_i)$ is the posterior probability. Similarly,

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{m} z_{ij}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^\top}{\sum_{i=1}^{m} z_{ij}} \tag{8}$$

Then, take derivatives with regard to each π_j and set it to zero

$$\sum_{i=1}^{m} \frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_l \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} + \lambda = \sum_{i=1}^{m} \frac{z_{ij}}{\pi_j} + \lambda \Longrightarrow \pi_j = -\frac{\sum_{i=1}^{m} z_{ij}}{\lambda}$$

With the constraint that $\sum_{j=1}^{k} \pi_j = -\sum_{i=1}^{m}\sum_{j=1}^{k} z_{ij}/\lambda = -m/\lambda = 1$, then $\lambda = -m$, and hence

$$\pi_j = \frac{\sum_{i=1}^{m} z_{ij}}{m} \tag{9}$$

It means that the mixing coefficient for the k-th component is given by the average posterior which that component takes for explaining the data points. Notes that the calculation above drops into a circle form: $\boldsymbol{\mu}, \boldsymbol{\Sigma} \to z_{ij} \to \boldsymbol{\mu}, \boldsymbol{\Sigma}$, hence we must do it in an iterative way, which is the EM algorithm for GMM:

- fix k, the number of Gaussian components;
- assign each sample to each components with equal probability, *i.e.*, $z_{ij} = \frac{1}{k}, j = 1, \cdots, k$ and $\pi_j = \frac{1}{k}, j = 1, \cdots, k$ also;
- M-step, solve $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ according to *Eq.*7 and *Eq.*8;
- E-step, solve z_{ij}, π_i according to *Eq.*9.

6 *v.s.* discriminant models

In generative approaches, it is assumed that the underlying distribution over the data has a specific parametric form and the goal is to estimate the parameters of the model. But in discriminative approaches, the goal is rather to learn an accurate predictor directly.

Of course, if we succeed in learning the underlying distribution accurately, prediction from the Bayes optimal classifier is reliable. The problem is that, it is usually more difficult to learn the underlying distribution than to learn an accurate predictor. This was phrased by Vladimir Vapnik:

"When solving a given problem, try to avoid a more general problem as an intermediate step."

However, in some situations, it is reasonable to adopt the generative models. Sometimes it is easier (computationally) to estimate the parameters of the model than to learn a discriminative predictor. Additionally, in some cases we do not have a specific task at hand but rather would like to use the data at a later time.

Modern generative models have another big goal, that is to 'generate' (sample from the underlying distribution) data like that in reality. The intuition behind this approach follows a famous quote from Richard Feynman:

"What I cannot create, I do not understand."

6.1 Naive Bayes to linear discriminant models

The usual assumption in Naive Bayes classifier is that each conditional probability $p(X = \boldsymbol{x}|Y = y)$ is a Gaussian distribution. Consider the binary classification task, denote the two conditional distribution as $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, we will predict $h_{\text{Bayes}}(\boldsymbol{x}) = 1$ iff.

$$\begin{aligned}
&\frac{p(Y=0)p(X=\boldsymbol{x}|Y=0)}{p(Y=1)p(X=\boldsymbol{x}|Y=1)} > 1 \\
\Longleftrightarrow\ &\log\frac{p(Y=0)}{p(Y=1)} + \log p(X=\boldsymbol{x}|Y=0) - \log p(X=\boldsymbol{x}|Y=1) > 0 \\
\Longleftrightarrow\ &\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1})\boldsymbol{x} + \underbrace{\frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0) + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} + \log\frac{p(Y=0)}{p(Y=1)}}_{\text{b}} > 0
\end{aligned}$$

which is a quadratic discriminant function. Further, if we assume that $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$, the classifier can be simplified to be a linear discriminant function $\boldsymbol{w} \cdot \boldsymbol{x} + b$, with $\boldsymbol{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}$ and $b = \frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0) + \log\frac{p(Y=0)}{p(Y=1)}$. If the prior probability is equal, namely $p(Y=0) = p(Y=1)$, the bias term can be further simplified.

7 Exercises and solutions

Ex1 **K-means** (see *UML Chapter 22.2, PRML Chapter 9.1*). K-means is a simple but important clustering algorithm. In fact, GMM is sometimes called *soft* K-means. As a hard version, K-means assigns the most probable cluster label to an example (*i.e.*, $z_i j = 1$ for one of $j \in 1, \cdots, k$ but 0 for others), and calculate the mean and covariance based on the in-cluster instead of global data. Formally, its procedure is as below,

- fix k, the number of clusters;
- randomly choose initial clustering centers $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k$
- repeat until convergence:
- 1. $\forall i \in \{1, \cdots, m\}$, determine $j = \arg\min_j d(\boldsymbol{x}_i - \boldsymbol{\mu}_j)$ and set $z_{ij} = 1$;
- 2. $\forall j \in \{1, \cdots, k\}$, update $\boldsymbol{\mu}_j = \frac{1}{\sum_{i=1}^m z_{ij}} \sum_{i=1, z_{ij}=1}^m \boldsymbol{x}_i$

in which $d(\cdot, \cdot)$ can be arbitrary distance function. Note that the step 1. corresponds to M-step of GMM, and step 2 corresponds to E-step. For GMM, the objective is to maximize likelihood, and for k-means, the objective can be viewed as minimizing the sum of in-cluster distance (if we choose the distance to be Euclidean distance, the loss is also called Sum of in-cluster Square Error, *a.k.a.*, SSE):

$$\min_{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k} \sum_{j=1}^k \sum_{i=1, z_{ij}=1}^m d(\boldsymbol{x}_i, \boldsymbol{\mu}_j)$$

Now, prove that: each iteration of the k-means algorithm does not increase the objective.

Solution:

Ex2 **Simplex of Dirichlet distribution** Because of the summation constraint, the distribution over the space of the $\{\mu_j\}$ is confined to a simplex of dimensionality $d - 1$.

Ex3 **Sequential estimation** (see *PRML Chapter 2.3.5*).

Ex4 **Sequential estimation under the perspective of Bayesian reasoning** (see *PRML Chapter 2.3.5*).

Ex5 **Unbiased estimation** (UML Ex24.1) θ_{ML}, in intrinsic, is a function of observed random variables, and hence has its expectation. If the expectation of an estimation is exactly the parameter in theory, we say that the estimation is unbiased. In the case of exponential family,

$$\mathbb{E}(\mu_{\mathrm{ML}}) = \mathbb{E}\left(\frac{\sum_{i=1}^m x_i}{m}\right) = \sum_{i=1}^m \frac{\mathbb{E}(x_i)}{m} = \mathbb{E}(x) = \mu$$

Hence, we say that the MLE for mean parameter is unbiased. Now, prove that the maximum likelihood estimator of the variance of a Gaussian variable is biased.

Solution:

$$\mathbb{E}(\boldsymbol{\Sigma}_{\mathrm{ML}}) = \sum_{i=1}^m \frac{\mathbb{E}((\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})^\top)}{m} = \sum_{i=1}^m \frac{\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i^\top) + \mathbb{E}(\boldsymbol{\mu}_{\mathrm{ML}} \boldsymbol{\mu}_{\mathrm{ML}}^\top) - 2\mathbb{E}(\boldsymbol{\mu}_{\mathrm{ML}} \boldsymbol{x}_i^\top)}{m}$$

Consider each term in the numerator, note that each pair of samples is independent,

$$\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i^\top) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

$$\mathbb{E}(\boldsymbol{\mu}_{\mathrm{ML}} \boldsymbol{\mu}_{\mathrm{ML}}^\top) = \frac{1}{m^2}\mathbb{E}\left(\sum_{i=1}^m \sum_{j=1}^m \boldsymbol{x}_i \boldsymbol{x}_j^\top\right) = \frac{1}{m^2}\mathbb{E}\left(\sum_{i=1}^m \sum_{j=1}^m (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^\top + 2\boldsymbol{\mu} \sum_{i=1}^m (\boldsymbol{x}_i - \boldsymbol{\mu})^\top + \sum_{i=1}^m \sum_{j=1}^m \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)$$

$$= \frac{1}{m^2}\left(\sum_{i=1}^m \mathbb{E}((\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top) + m^2 \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) = \frac{\boldsymbol{\Sigma}}{m} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

$$\mathbb{E}(\boldsymbol{\mu}_{\mathrm{ML}} \boldsymbol{x}_i^\top) = \mathbb{E}\left(\frac{1}{m}\sum_{j=1}^m \boldsymbol{x}_j \boldsymbol{x}_i^\top\right) = \frac{1}{m}\mathbb{E}\left(\sum_{j=1}^m (\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top + 2\boldsymbol{\mu} \sum_{j=1}^m (\boldsymbol{x}_j - \boldsymbol{\mu})^\top + \sum_{j=1}^m \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) = \frac{\boldsymbol{\Sigma}}{m} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

Hence, $\mathbb{E}(\boldsymbol{\Sigma}_{\mathrm{ML}}) = \frac{m-1}{m}\boldsymbol{\Sigma}$ which is biased.

Ex6 **The connection between smoothing and regularized MLE** (UML Ex24.2) Consider the following regularized loss minimization for parameter estimation in the case of Bernoulli distribution:

$$\min \frac{1}{m} \sum_{i=1}^{m} -\log \mathcal{P}_{\boldsymbol{\mu}}(\boldsymbol{x}_i) + \frac{1}{m}(\log(1/\mu) + \log(1/(1-\mu)))$$

6.1 Show that the preceding objective is equivalent to the usual empirical error had we added two pseudo-examples to the training set.

6.2 Derive a high probability bound on $|\mu' - \mu|$, and use this to bound the true risk.

Solution:

6.1 The regularized loss can be written as

$$-\frac{1}{m} \sum_{i=1}^{m} x_i \log \mu + (1 - x_i) \log(1-\mu) - \frac{1}{m}(\log(\mu) + \log(1-\mu))$$

Take derivatives with regard to μ and set it to zero leads to $\mu' = \frac{1+\sum_{i=1}^{m} x_i}{m+2}$. It's equivalent to adding two pseudo-examples $\{0, 1\}$ into the training set, which is called 'add-1' smoothing.

6.2 Using triangle inequality,

$$|\mu' - \mu| = |\mu' - \mathbb{E}(\mu') + \mathbb{E}(\mu') - \mu| \leq |\mu' - \mathbb{E}(\mu')| + |\mathbb{E}(\mu') - \mu|$$

Since $\mathbb{E}(\mu') = \frac{1+m\mu}{m+2}$, we have that $|\mathbb{E}(\mu') - \mu| \leq \frac{1}{m+2}$, and $|\mu' - \mathbb{E}(\mu')| = \frac{m}{m+2} |\frac{1}{m} \sum_{i=1}^{m} x_i - \mu|$. Following Hoeffding's inequality, for any $\epsilon > 0$,

$$P\left(|\mu' - \mu| \geq \frac{1}{m+2} + \epsilon\right) \leq 2 \exp\left(-2m\epsilon^2\right)$$

Chapter 4. Linear models for classification and regression, penalization
Chapter 5. Decision stumps, ensemble learning, Bayes PAC
Chapter 6. Perceptron, MLP, deep learning, Generalization bounds on deep learning.