Chapter *3* Generative Models

Siheng Zhang
zhangsiheng@cvte.com

October 30, 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 Density estimation

To apply the Bayesian decision principle, we should know the probability distribution. In fact, machine learning can be treated as '***fitting the underlying distribution***' (see 4). There are two classes of methods for estimation, parametric and non-parametric methods.

2.1 Parametric method: maximum likelihood

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, that is

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

Following is the examples:

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

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

The derivative of the log likelihood with respect to μ is given by

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

which leads to $\mu_{\text{ML}} = \frac{1}{m} \sum_{i=1}^{m} x_i$.

remark1: θ_{ML}, in intrinsic, is a function of observed random variables, and hence we can calculate its expectation. If the expectation of an estimation is exactly the parameter in theory, we say that the estimation is unbiased. In this example,

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

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

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{j=1}^{d} \mu_j^{x_j}$$

in which $\sum_{j=1}^{d} \mu_j = 1, \mu_j \geq 0$.

The corresponding log likelihood function is given by

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

To maximize the log likelihood function with respect to μ_j must take account of the constraint that $\sum_{j=1}^{d} \mu_j = 1$. Using Lagrange multiplier λ, we should maximize

$$L = \sum_{i=1}^{m}\sum_{j=1}^{d} x_{ij} \log \mu_j + \lambda\left(\sum_{j=1}^{d} \mu_j - 1\right)$$

Take derivative with regard to μ_j

$$\frac{\partial L}{\partial \mu_j} = \sum_{i=1}^{m} \frac{x_{ij}}{\mu_j} + \lambda$$

which leads to $\mu_{j,\mathrm{ML}} = -\sum_{i=1}^{m} x_{ij}/\lambda$. Recall that $\sum_{j=1}^{d} \mu_j = -\sum_{j=1}^{d}\sum_{i=1}^{m} x_{ij}/\lambda = -m/\lambda = 1$, λ is set to $-m$, and hence,

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \tag{3}$$

which is also unbiased.

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

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{4}$$

The log likelihood function is given by

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

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

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

and setting it to zero leads to: $\boldsymbol{\mu}_{\mathrm{ML}} = \sum_{i=1}^{m} \boldsymbol{x}_i/m$.

remark2: 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 $\boldsymbol{x}^{\top}\boldsymbol{Ax} = tr[\boldsymbol{x}^{\top}\boldsymbol{Ax}] = tr[\boldsymbol{xx}^{\top}\boldsymbol{A}]$;

- $\frac{\partial tr[\boldsymbol{AB}]}{\partial \boldsymbol{A}} = \boldsymbol{B}^\top$;
- $\frac{\partial \log |\boldsymbol{A}|}{\partial \boldsymbol{A}} = (\boldsymbol{A}^{-1})^\top$;
- $\frac{\partial tr(\boldsymbol{AX}^{-1}\boldsymbol{B})}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1}\boldsymbol{BAX}^{-1})^\top$

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:

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \sum_{i=1}^{m} \frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})^\top}{m} \tag{5}$$

remark3: Again, we consider the property of MLE result. The estimation of $\boldsymbol{\mu}$ is unbiased. However, the estimation of $\boldsymbol{\Sigma}$ is biased,

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

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

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

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

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

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\{\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\eta})\} \tag{6}$$

remark4: Bernoulli distribution is a member in this family,

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

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

remark5: 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

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{j=1}^{d}\mu_j^{x_j} = \exp\left\{\sum_{j=1}^{d}x_j\ln\mu_j\right\} = \exp\left\{\sum_{j=1}^{d-1}x_j\ln\mu_j + \left(1 - \sum_{j=1}^{d-1}x_k\right)\ln\left(1 - \sum_{j=1}^{d-1}\mu_k\right)\right\}$$

Compare with the general form shows that

remark6: Gaussian distribution is a member in this family.

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

Compare with the general form shows that

Now consider the problem of estimating the parameter vector $\boldsymbol{\mu}$ in the general exponential family distribution.

2.2 Expectation maximization: maximum likelihood 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:

$$L(S;\theta) = \prod_{i=1}^{m}\sum_{j=1}^{k} p_\theta(x_i, y_i) = \prod_{i=1}^{m}\sum_{j=1}^{k} p_\theta(x_i|y_i)p_\theta(y_i)$$

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

$$\log L(S;\theta) = \sum_{i=1}^{m} \log p_\theta(x_i|y_i)p_\theta(y_i) \tag{7}$$

GMM (Gaussian mixture models) is a typical example.

2.3 Non-parametric methods

3 Bayesian Reasoning

4 *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 the real world. The intuition behind this approach follows a famous quote from Richard Feynman:

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

4.1 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.

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

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.

5 Exercises and solutions

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