Machine learning - Generative Learning Algorithms

Yonghoon Dong

March 8, 2024

Discriminative learning / Generative learning

1. Discriminative learning algorithm : try to learn $p(y|x)$ directly
2. Generative learning algorithm : try to model $p(x|y)$ and $p(y)$ and use Bayes rule to derive $p(y|x)$

Examples of generative learning

We will learn two generative learning algorithm:

1. Gaussian Discriminant Analysis : continuous random variable case
2. Naive Bayes : discrete random variable case

Gaussian Discriminant Analysis

For a **classification** problem in which the input features x are continuous random variables, **Gaussian Discriminant Analysis** can be used. The model is

$$y \sim \text{Ber}(\phi)$$
$$x|y=0 \sim \mathcal{N}(\mu_0, \Sigma)$$
$$x|y=1 \sim \mathcal{N}(\mu_1, \Sigma)$$

Alternatively,

$$p(y) = \phi^y (1-\phi)^{1-y}$$
$$p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)$$
$$p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_1)\right)$$

Gaussian Discriminant Analysis

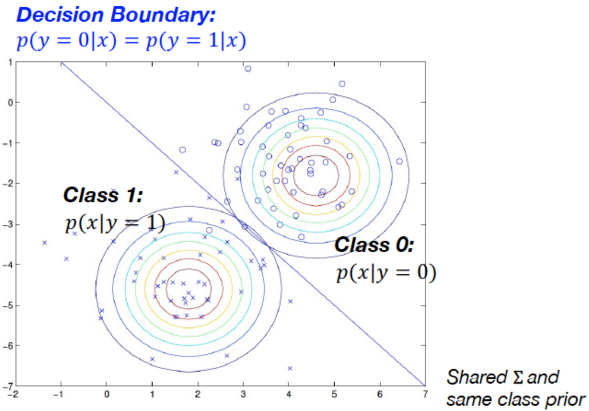

Figure: Illustration of the Gaussian Discriminant Analysis when covariance matrices are same

Gaussian Discriminant Analysis

Define a loss function by using negative log probability

$$\begin{aligned}\ell(y, f(x, \phi, \mu_0, \mu_1, \Sigma) &= -\log p(x, y; \mu_0, \mu_1, \Sigma)\\ &= -\log p(x|y; \mu_0, \mu_1, \Sigma)p(y; \phi)\end{aligned}$$

Note that we use **joint likelihood**

Gaussian Discriminant Analysis

Therefore, our task is to find optimal $\phi, \mu_0, \mu_1, \Sigma$ such that

$$\begin{aligned}
\arg\min_{\phi,\mu_0,\mu_1,\Sigma} \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, f(x_n, \phi, \mu_0, \mu_1, \Sigma)) &= -\frac{1}{N} \sum_{n=1}^{N} \log p(x_n, y_n; \mu_0, \mu_1, \Sigma) \\
&= -\frac{1}{N} \sum_{n=1}^{N} \log p(x_n | y_n; \mu_0, \mu_1, \Sigma) p(y_n; \mu_0, \mu_1, \Sigma)
\end{aligned}$$

Gaussian Discriminant Analysis

By maximizing the likelihood (i.e. minimizing the empirical risk) with respect to the parameters, we find the optimal parameters

$$\phi = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{y_n = 1\} : \text{allocation ratio of class 1}$$

$$\mu_0 = \frac{\sum_{n=1}^{N} \mathbf{1}\{y_n = 0\} x_n}{\sum_{n=1}^{N} \mathbf{1}\{y_n = 0\}} : \text{average value of class 0}$$

$$\mu_1 = \frac{\sum_{n=1}^{N} \mathbf{1}\{y_n = 1\} x_n}{\sum_{n=1}^{N} \mathbf{1}\{y_n = 1\}} : \text{average value of class 1}$$

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{y_n})(x_n - \mu_{y_n})^T : \text{average covariance matrix}$$

GDA and logistic regression

If we view the quantity $p(y = 1|x; \phi, \mu_0, \mu_1, \Sigma)$ as a function of x, we'll find that it can be expressed in the form

$$p(y = 1|x; \phi, \Sigma, \mu_0, \mu_1) = \frac{1}{1 + \exp(-\theta^T x)}$$

where θ is some appropriate function of $\phi, \Sigma, \mu_0, \mu_1$. This is exactly the form that logistic regression.

GDA and logistic regression

Which is better?

1. If $p(x|y)$ is multivariate Gaussian with shared Σ, then $p(y|x)$ necessarily follows a logistic function.
2. $p(y|x)$ being a logistic function doesn't imply $p(x|y)$ is multivariate Gaussian.
3. GDA makes **stronger** assumptions about the data than does logistic regression

When $p(x|y)$ is Gaussian, then GDA is efficient. In contrast, by making weaker assumptions, logistic regression is more robust and less sensitive to incorrect modeling assumptions