2. Probabilistic Generative Model

In "1, Discriminant Functions" we are using the simplest model.
Let's illustrate all 3 approaches, and discuss the differences.

Three approaches to classification

a) Find $f(x)$, which is called discriminant function, to map input x directly into a class label.

This approach doesn't care anything related to probability.

Complexity *

b) Determine the posterior class probability $p(C_k|x)$, then use decision theorem to do the rest.

This approach needs to compute the posterior, so a little bit complexer than a). $P(C_k|x)$ is called the discriminant model.

Complexity **

c) Find out class-conditional density $P(x|C_k)$

Then use Bayes' theorem to compute posterior class probability $P(C_k|x)$

Or, equivalently, compute the joint distribution $p(x, C_k)$, then reduce to $p(C_k|x)$.

Finally use decision theorem to classify x.

Because this approach explicitly or implicitly model the distribution of input and output, this is known as generative models.

Complexity *****

For generative model approach, nearly every thing can be computed out. So, if we sample $p(x, C_k)$, we can even generate synthetic data in the input space.

Because we need to compute so many "irralevent" stuff, and the dimension of x is usually large, the complexity is usually really high.

Let's consider the case of two classes. The posterior probability for class C_1 can be written as

$$p(C_1 | x) = \frac{p(x|C_1) p(C_1)}{p(x|C_1) \cdot p(C_1) + p(x|C_2) p(C_2)}$$

$$= \frac{1}{1+\exp(-\alpha)} = \sigma(\alpha)$$

$$\alpha = \ln \frac{p(x|C_1) p(C_1)}{p(x|C_2) p(C_2)}$$

$\sigma(\alpha)$ is the logistic sigmoid function defined by

$$\sigma(\alpha) = 1/[1+\exp(-\alpha)]$$

Let's analyze sigmoid first

$$\sigma(-a) = 1 - \sigma(a), \quad a = \ln\left(\frac{\sigma}{1-\sigma}\right)$$

1. Continuous Input.

Let's assume that every class's pdf is Gaussian, and all classes share the same covariance matrix Σ. Then, the density for class C_k is given by

$$p(x|C_k) = \mathcal{N}(x|\mu_k, \Sigma)$$

Now, use the sigmoid function version.

$$P(C_1|x) = \sigma(W^T x + b_0)$$

where $W = \Sigma^{-1}(\mu_1 - \mu_2)$

$$b_0 = -\frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \ln\frac{P(C_1)}{P(C_2)}$$

We can see that, the quadratic term relates x has been cancelled. Then, we find that the argument of logistic sigmoid is linear to x.

2. Logistic Regression

Also for the two-class classification, we know from previous derivation that

$$p(C_1|\phi(x)) = \sigma(\tilde{w}^T \tilde{\phi}(x))$$

$$p(C_2|\phi(x)) = 1 - p(C_1|\phi(x))$$

In statistics, this method is known as logistic regression, even though it's a classification model.

For a dataset $\{\phi_n, t_n\}$, where $t_n \in \{0, 1\}$, and $\phi_n = \phi(x_n)$

The likelihood function can be written as

$$p(\mathbf{t} | w) = \prod_{n=1}^{N} p(C_1 | \phi_n)^{t_n} \underbrace{(1 - p(C_1 | \phi_n))}_{= p(C_2 | \phi_n)}^{1-t_n}$$

As usual, take the negative logarithm of the likehood, which is the cross-entropy error

$$E(w) = -\ln p(\mathbf{t} | w) = -\sum_{n=1}^{N} \{ t_n \ln p(C_1 | \phi_n) + (1-t_n) \ln (1 - p(C_1 | \phi_n)) \}$$

Now, plug in our sigmoid function and linear model.

$$p(C_1 | \phi_n) = \sigma(w^T \phi_n)$$

Then

$$\nabla_w E(w) = \sum_{n=1}^{N} (\sigma(w^T \phi_n) - t_n) \phi_n$$