CS3390: Foundations of Machine Learning **2023**

Lecture 7: 12 September 2023

Instructor: P. K. Srijith *Scribe: Gautam Singh*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

7.1 Probabilistic Generative Models

So far, we have considered the following classification methods.

1. Using posterior class probabilities, called *discriminative models*.
2. Find a function f, called a discriminant function, which maps each input $\mathbf{x}$ directly onto a class label. Probabilities play no role here.

7.1.1 Generative Classifiers

Here, we first solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|y)$. Also, we require the prior class probabilities $p(y)$. Then, we use Bayes' Theorem as follows.

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)\, p(y)}{p(\mathbf{x})} \tag{7.1}$$

Approaches that implicitly or explicitly model the distribution of inputs and outputs are called **generative models**, because by sampling from them we can generate synthetic data points in the input space.

Approaches that model the posterior probabilities directly are called **discriminative models**.

7.1.1.1 Binary Classification

For two classes, we have

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)\, p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)\, p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)\, p(\mathcal{C}_2)} \tag{7.2}$$

$$= \frac{1}{1 + \exp(-a)} \tag{7.3}$$

where

$$a \triangleq \ln \frac{p(\mathbf{x}|\mathcal{C}_1)\, p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)\, p(\mathcal{C}_2)}. \tag{7.4}$$

7.1.1.2 Multiclass Classification

If there are k classes,

$$p\left(\mathcal{C}_k|\mathbf{x}\right) = \frac{p\left(\mathbf{x}|\mathcal{C}_k\right) p\left(\mathcal{C}_k\right)}{\sum_i p\left(\mathbf{x}|\mathcal{C}_k\right) p\left(\mathcal{C}_k\right)} \tag{7.5}$$

$$= \frac{\exp\left(a_k\right)}{\sum_j \exp a_j} \tag{7.6}$$

where

$$a_k \triangleq \ln p\left(\mathbf{x}|\mathcal{C} - k\right) p\left(\mathcal{C}_k\right) \tag{7.7}$$

Hence,

$$p\left(\mathbf{x}|\mathcal{C}_k\right) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\mathbf{\Sigma}\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_\mathbf{k}{}^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_\mathbf{k}\right)\right)\right\} \tag{7.8}$$

and also,

$$\mathbf{w_k} = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_\mathbf{k} \tag{7.9}$$

7.2 Discriminant Analysis: Parameter Estimation

We model joint probability of observing input and output.

$$p\left(\mathbf{x_n}, \mathcal{C}_1\right) = p\left(\mathcal{C}_1\right) p\left(\mathbf{x_n}|\mathcal{C}_1\right) = \pi\mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{1}, \mathbf{Sigma}\right) \tag{7.10}$$

$$p\left(\mathbf{x_n}, \mathcal{C}_2\right) = p\left(\mathcal{C}_2\right) p\left(\mathbf{x_n}|\mathcal{C}_2\right) = (1 - \pi)\,\mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{2}, \mathbf{Sigma}\right) \tag{7.11}$$

Then,

$$p\left(\mathbf{t}|\pi, \boldsymbol{\mu}_\mathbf{1}, \boldsymbol{\mu}_\mathbf{2}, \mathbf{\Sigma}\right) = \prod_{i=1}^{N} \left[\pi\mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{1}, \mathbf{\Sigma}\right)\right]^{t_n} \left[(1 - \pi)\,\mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{2}, \mathbf{\Sigma}\right)\right]^{1 - t_n} \tag{7.12}$$