Parametric classification and regression

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 5

parametric classification and regression in 1D

Classification

K classes $(C_1, \cdots, C_K)$

Recall: Bayes decision rule

$$k = \underset{i}{\operatorname{argmax}}\ P(C_i \mid x)$$

Equivalently, we can define a discriminant function

$$\bar{g}_i(x) = p(x \mid C_i)\, P(C_i) \qquad k = \underset{i}{\operatorname{argmax}}\ g_i(x)$$

Also equivalently, we can choose to look at

$$\boxed{g_i(x) = \log p(x \mid C_i) + \log P(C_i)}$$

Consider a parametric model for $p(x \mid C_i)$

Assume $p(x \mid C_i) = N(x \mid \mu_i, \sigma_i^2)$, $i = 1, \cdots, K$.

$$= \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

Then  becomes

$$\boxed{g_i(x) = -\frac{1}{2}\log(2\pi) - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)}$$

Assume we have a sample $\mathcal{X} = \{ \mathbb{X}_n \}_{n=1}^{N}$

where each $\mathbb{X}_n = (x_n , t_n)$,

with $x_n \in \mathbb{R}$

$t_n \in \{0, 1\}^K$ one-hot

such that

$$t_{ni} = \begin{cases} 1 & \text{if } \mathbb{X}_n \in C_i \\ 0 & \text{if } \mathbb{X}_n \notin C_i \end{cases}$$

Suppose we use MLE for the parameters.

For each class i,

$$\left.\begin{aligned} \hat{\mu}_i &= \frac{\sum_{n=1}^{N} x_n t_{ni}}{\sum_{n=1}^{N} t_{ni}} \\ \hat{\sigma}_i &= \frac{\sum_{n=1}^{N} (x_n - \hat{\mu}_i)^2 t_{ni}}{\sum_{n=1}^{N} t_{ni}} \\ \hat{P}(C_i) &= \frac{\sum_{n=1}^{N} t_{ni}}{N} \end{aligned}\right\} \text{plug these in}$$

$$\hat{g}_i(x) = -\frac{1}{2} \log (2\pi) - \log \hat{\sigma}_i - \frac{(x - \hat{\mu}_i)^2}{2 \hat{\sigma}_i^2} + \log \hat{P}(C_i)$$

Now that we have

$$\hat{g}_i(x) = -\frac{1}{2}\log(2\pi) - \log\hat{\sigma}_i - \frac{(x-\hat{\mu}_i)^2}{2\hat{\sigma}_i^2} + \log\hat{P}(C_i)$$

If we further assume

i. The priors are equal: $\hat{P}(C_1) = \cdots = \hat{P}(C_K)$

ii. the variances are equal: $\hat{\sigma}_1 = \cdots = \hat{\sigma}_K$.

Then

$$k = \arg\max_i \hat{g}_i(x)$$

$$= \arg\max_i -(x-\hat{\mu}_i)^2$$

$$= \arg\min_i (x-\hat{\mu}_i)^2$$

Regression

$$y = f(x) + \varepsilon$$

f is unknown, estimated using $g(x|\theta)$

Assume $\quad p(\varepsilon) = \mathcal{N}(\varepsilon \mid 0, \sigma^2)$

Then $p(y|x) = \mathcal{N}(y \mid g(x|\theta), \sigma^2)$

Assume we have a sample $\mathcal{X} = \{\mathbb{X}_n\}_{n=1}^N$,

where each $\mathbb{X}_n = (x_n, y_n) \overset{\text{i.i.d.}}{\sim} p(x, y)$

Note that $p(x,y) = p(y|x)\, p(x)$

Therefore,

$$\log(\mathcal{X}|\theta) = \log \prod_{n=1}^{N} p(x_n, y_n)$$

$$= \log \prod_{n=1}^{N} p(y_n|x_n)\, p(x_n)$$

$$= \underbrace{\log \prod_{n=1}^{N} p(y_n|x_n)}_{\text{depends on } \theta} + \log \prod_{n=1}^{N} p(x_n)$$

Therefore,

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log \prod_{n=1}^{N} p(y_n | x_n)$$

$$= \arg\max_{\theta} \log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y_n - g(x_n|\theta))^2}{2\sigma^2}\right)$$

$$= \arg\max_{\theta} \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y_n - g(x_n|\theta))^2}{2\sigma^2}\right)$$

$$= \arg\max_{\theta} -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - g(x_n|\theta))^2$$

$$= \arg\min_{\theta} \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y_n - g(x_n|\theta))^2}_{E(\theta|\mathcal{X})}$$

error function

a.k.a. "least squares estimation".

e.g. linear regression

$$g(x | w_0, w_1) = w_0 + w_1 x .$$

The corresponding error function is

$$E(w_0, w_1 | \mathcal{X}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - w_0 - w_1 x_n)^2$$

Now that

$$E(w_0, w_1 \mid x) = \frac{1}{2} \sum_{n=1}^{N} (y_n - w_0 - w_1 x_n)^2$$

$$\begin{cases} \dfrac{\partial E}{\partial w_0} = -\displaystyle\sum_{n=1}^{N} (y_n - w_0 - w_1 x_n) \overset{\text{set}}{=\!=} 0 \\ \dfrac{\partial E}{\partial w_1} = -\displaystyle\sum_{n=1}^{N} (y_n - w_0 - w_1 x_n)\, x_n \overset{\text{set}}{=\!=} 0 \end{cases}$$

That is,

$$\begin{cases} N w_0 + \left(\displaystyle\sum_{n=1}^{N} x_n\right) w_1 = \displaystyle\sum_{n=1}^{N} y_n \\ \left(\displaystyle\sum_{n=1}^{N} x_n\right) w_0 + \left(\displaystyle\sum_{n=1}^{N} x_n^2\right) w_1 = \displaystyle\sum_{n=1}^{N} y_n x_n \end{cases}$$

Write $A = \begin{pmatrix} N & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 \end{pmatrix}$, $w = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$, $r = \begin{pmatrix} \sum_{n=1}^{N} y_n \\ \sum_{n=1}^{N} y_n x_n \end{pmatrix}$

Need to solve:

$$A w = r.$$

$$|A| = N \sum_{n=1}^{N} x_n^2 - \left(\sum_{n=1}^{N} x_n\right)^2 \geq 0$$

by Cauchy-Schwartz inequality,

and $|A| = 0 \iff$ All x_n's are equal.

When $|A| > 0$, $\quad w = A^{-1} r.$

Gaussian density in high dimension

- In 1-dimension,

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- In D-dimension,

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gaussian density in high dimension

- From $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$, it is clear that the Gaussian density depends on $\boldsymbol{x}$ through

$$\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

- Δ is called the Mahalanobis distance from $\boldsymbol{\mu}$ to $\boldsymbol{x}$.
- This reduces to the Euclidean distance if $\boldsymbol{\Sigma} = \mathbf{I}$.

multivariate classification

Recall: $g_i(x) = \log p(x|C_i) + \log P(C_i)$

Now suppose $x \in \mathbb{R}^D$. Assume

$$p(x|C_i) = \mathcal{N}(x|\mu_i, \Sigma_i)$$

$$= \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right)$$

$$g_i(x) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \log P(C_i)$$

Consider a sample $\mathcal{X} = \{\underline{x}_n\}_{n=1}^N = \{(x_n, t_n)\}_{n=1}^N$

one-hot

The MLE satisfies

$$\begin{cases} \hat{\mu}_i = \dfrac{\sum_{n=1}^N t_{ni}\, x_n}{\sum_{n=1}^N t_{ni}} =: m_i \\[2ex] \hat{\Sigma}_i = \dfrac{\sum_{n=1}^N t_{ni}\,(x_n-\hat{\mu}_i)(x_n-\hat{\mu}_i)^T}{\sum_{n=1}^N t_{ni}} =: S_i \quad \text{(abuse of notation)} \\[2ex] \hat{P}(C_i) = \dfrac{\sum_{n=1}^N t_{ni}}{N} \end{cases}$$

$$\hat{g}_i(x) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|S_i| - \frac{1}{2}(x-m_i)^T S_i^{-1}(x-m_i)$$
$$+ \log \hat{P}(C_i)$$

If we disregard the constant, we can also look at

$$\check{g}_i(x) = -\frac{1}{2}\log|S_i| - \frac{1}{2}\left(x^T S_i^{-1} x - 2x^T S_i^{-1} m_i + m_i^T S_i^{-1} m_i\right)$$
$$+ \log \hat{P}(C_i)$$

$$= -\frac{1}{2} x^T S_i^{-1} x + x^T S_i^{-1} m_i$$
$$- \frac{1}{2}\log|S_i| - \frac{1}{2} m_i^T S_i^{-1} m_i + \log \hat{P}(C_i)$$

$$= x^T W_i x + w_i^T x + w_{i0}$$

Questions?

Reference

- *Parametric classification and regression:*
  - *[Al] Ch.4.5*
- *Multivariate methods:*
  - *[Al] Ch.5.1-5.5*
  - *[Bi] Ch.2.3*