Probabilistic Machine Learning

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Exercise 9

Summer Term 2023

Full Name	Matriculation No
Batuhan Özcömlekci	6300476
Aakarsh Nair	6546577

Eberhards Karls University of Tübingen, July 3, 2023

1. Theory Question: In the lecture we encountered the plot reproduced here, showing a structured classification problem. It was pointed out that the most salient feature of this problem is that each class seems to have its own generative distribution.

i.e the blue points have a different distribution than the green ones. (although both distributions happen to overlap).

Such situations indicate an anti-causal relationship. ("the label causes the inputs")

and it might seem that discriminative modeling paradigm adopted in the lecture is not appropriate.

The goal of this week's theory is to realise that the situation is a bit more subtle than that.

For the purpose of this exercise, we will assume that there are two classes C_1 and C_2, defining probability distributions

$p(x|C_1)$, $p(x|C_2)$ over the inputs, and

classes are drawn with the probability

$$p(C) = [p(C_1), p(C_2) = 1 - p(C_1)]$$

(a) Given a new input x, how would you compute the posterior $p(C_1|x)$? Show that it can be written as a logistic function.

Ans: $p(C_1|x) = \dfrac{1}{1+e^{-a(x)}}$ where $a(x) = \ln \dfrac{p(C_1|x)}{p(C_2|x)}$

Consider the RHS

$$\Rightarrow \frac{1}{1+e^{-a(x)}} = \frac{1}{1+e^{-\ln\left(\frac{p(C_1|x)}{p(C_2|x)}\right)}} = \frac{1}{1+\frac{p(C_2|x)}{p(C_1|x)}} = \frac{p(C_1|x)}{p(C_1|x)+p(C_2|x)}$$

using Bayes rule

$$\Rightarrow \frac{p(C_1|x)}{\frac{p(x|C_1)\,p(C_1)}{p(x)} + \frac{p(x|C_2)\,p(C_2)}{p(x)}}$$

$$\Rightarrow p(C_1|x)\left[\frac{p(x)}{p(x|C_1)\,p(C_1) + p(x|C_2)\,p(C_2)}\right]$$

From law of total probability $p(x) = p(x|C_1)\,p(C_1) + p(x|C_2)\,p(C_2)$

$$\Rightarrow p(C_1|x)\left[\frac{p(x)}{p(x)}\right] = p(C_1|x)$$

1(b). This observation suggests that when we perform logistic regression to learn the function $a(x)$, we actually indirectly learn the class distributions $p(x|C_1)$ and $p(x|C_2)$. It is interesting to consider how different regression models for $a(x)$ relate to different assumptions about class distributions. Assume that both classes are draws from the same exponential family, with different parameters.

$$p(x|C_k) = h(x) \exp(\phi(x)^T \omega_k - \log Z(\omega_k))$$

Show that this implies a <u>linear</u> model for $a(x)$

$$a(x) = \phi(x)^T \vec{\theta} + \theta_0$$

What are the parameters $\vec{\theta}$ and θ_0 of this model in terms of the parameters ω of the class distributions?

Answer:

① Let $p(x|C_1) = h(x) \exp(\phi(x)^T \omega_1 - \log Z(\omega_1))$
and $p(x|C_2) = h(x) \exp(\phi(x)^T \omega_2 - \log Z(\omega_2))$

then we have $\dfrac{p(C_1|x)}{p(C_2|x)} = \dfrac{p(x|C_1)\, p(C_1) / p(x)}{p(x|C_2)\, p(C_2) / p(x)} = \dfrac{p(x|C_1)}{p(x|C_2)} \dfrac{p(C_1)}{p(C_2)}$

$$\Rightarrow \frac{\cancel{h(x)} \exp(\phi^T(x)\, \omega_1 - \log Z(\omega_1))}{\cancel{h(x)} \exp(\phi^T(x)\, \omega_2 - \log Z(\omega_2)} \quad \frac{p(C_1)}{p(C_2)}$$

$$\Rightarrow \exp\left[\phi^T(x)(\omega_1 - \omega_2) - \log Z(\omega_1) + \log Z(\omega_2)\right] \quad \frac{p(C_1)}{p(C_2)}$$

$$\Rightarrow \exp\left[\phi^T(x)(\omega_1 - \omega_2) - \ln \frac{Z(\omega_1)}{Z(\omega_2)} + \ln \frac{p(C_1)}{p(C_2)}\right]$$

We have $a(x) = \ln \dfrac{p(C_1|x)}{p(C_2|x)}$

Thus

$$a(x) = \ln \left[\exp \left[\phi^T(x)(\omega_1 - \omega_2) - \ln \frac{Z(\omega_1)}{Z(\omega_2)} + \ln \frac{P(C_1)}{P(C_2)} \right] \right]$$

$$a(x) = \phi^T(x)(\omega_1 - \omega_2) + \left[\ln \frac{P(C_1)}{P(C_2)} - \ln \frac{Z(\omega_1)}{Z(\omega_2)} \right]$$

thus $\vec{\theta} = \omega_1 - \omega_2$

$$\theta_0 = \ln \frac{P(C_1)}{P(C_2)} - \ln \frac{Z(\omega_1)}{Z(\omega_2)}$$

and $a(x) = \phi^T(x)\,\vec{\theta} + \theta_0$ is a linear model

1.c

$$\theta_0 = \ln \frac{p(C_1)}{1-p(C_1)} - \ln \frac{Z(w_1)}{Z(w_2)}$$

$$\vec{\theta} = w_1 - w_2$$

$$\alpha(x) = \phi^T(x)\vec{\theta} + \theta_0 \quad \text{linear model} = \ln \frac{p(C_1|x)}{p(C_2|x)}$$

$$\alpha(x) = \ln \frac{p(x|C_1)\,p(C_1)}{p(x|C_2)\,p(C_2)} = \ln \frac{\mathcal{N}(x;\mu_1,\sigma_1^2)\,p(C_1)}{\mathcal{N}(x;\mu_2,\sigma_2^2)\,p(C_2)}$$

$$= \ln \left(\frac{\frac{1}{\sqrt{2\pi}\sigma_1} \exp[-(x-\mu_1)^2/2\sigma_1^2] \quad p(C_1)}{\frac{1}{\sqrt{2\pi}\sigma_2} \exp[-(x-\mu_2)^2/2\sigma_2^2] \quad p(C_2)} \right)$$

$$= \ln \left(\frac{\sigma_2}{\sigma_1} \frac{\exp[-(x-\mu_1)^2/2\sigma_1^2]}{\exp[-(x-\mu_2)^2/2\sigma_2^2]} \frac{p(C_1)}{p(C_2)} \right)$$

$$= -(x-\mu_1)^2/2\sigma_1^2 + (x-\mu_2)^2/2\sigma_2^2 + \ln \frac{\sigma_2}{\sigma_1} + \ln \frac{p(C_1)}{p(C_2)}$$

$$= \frac{-x^2+2\mu_1 x-\mu_1^2}{2\sigma_1^2} + \frac{x^2-2\mu_2 x+\mu_2^2}{2\sigma_2^2} + \ln \frac{\sigma_2}{\sigma_1} + \ln \frac{p(C_1)}{p(C_2)}$$

$$= \left(-\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right)x^2 + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)x - \left(\frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{2\sigma_2^2}\right) + \ln \frac{\sigma_2}{\sigma_1} + \ln \frac{p(C_1)}{p(C_2)}$$

$$= \underbrace{\begin{bmatrix} x & -\frac{x^2}{2} \end{bmatrix}}_{\phi^T(x)} \underbrace{\begin{bmatrix} \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \\ \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \end{bmatrix}}_{\vec{\theta}} \underbrace{- \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) + \ln \frac{\sigma_2}{\sigma_1} + \ln \frac{p(C_1)}{1-p(C_1)}}_{\theta_0}$$

We can't recover $\mu_1, \mu_2, \sigma_1, \sigma_2, p(C_1)$ given $\vec{\theta}, \theta_0$ since there are some free variables (and infinitely many solutions).

If we set $p(C_1)=p(C_2)=\frac{1}{2}$, $\sigma_1=\sigma_2=1$, then $\vec{\theta}$ and θ_0 becomes

$$\vec{\theta} = \begin{bmatrix} \mu_1-\mu_2 \\ 0 \end{bmatrix}, \quad \theta_0 = -\frac{\mu_1^2-\mu_2^2}{2} = -\frac{(\mu_1-\mu_2)(\mu_1+\mu_2)}{2} \Rightarrow \mu_1+\mu_2 = -\frac{2\theta_0}{(\vec{\theta})_1}$$

Finally, we can recover $\mu_1 = \dfrac{(\vec{\theta})_1 - \frac{2\theta_0}{(\vec{\theta})_1}}{2}$, $\mu_2 = \dfrac{(\vec{\theta})_1 + \frac{2\theta_0}{(\vec{\theta})_1}}{2}$ where $(\vec{\theta})_1 = \mu_1-\mu_2$