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.

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

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

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 distri-butions $p(x|C_1)$ 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.

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

Show that this implies a linear 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 $p(x|C_1) = h(x) \exp\left(\phi(x)^T \omega_1 - \log Z(\omega_1)\right)$
and $p(x|C_2) = h(x) \exp\left(\phi(x)^T \omega_2 - \log Z(\omega_2)\right)$

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

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

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

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

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

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) Does this mean we can turn any discriminative model (predicting class labels from inputs, $p(C_k|x)$ into a generative one (predicting inputs from class labels, $p(x|C_k)$ by "going backwards" through the model? Unfortunately, the answer is no, because continuous distributions for inputs are more complex than binary distributions.

To make this clear, consider the special case: Assume that the class distributions are both univariate Gaussians with different means and variances:

$$p(x|C_k) = \mathcal{N}(x; \mu_k, \sigma_k^2).$$

From your answer above, construct an explicit form of θ, θ_0 (that's two real numbers) as a function of $\mu_1, \mu_2, \sigma_1, \sigma_2, p(C_1)$.

Assume someone has performed logistic regression and given you the parameters θ, θ_0.

1. Can you recover the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, p(C_1)$?
2. Could you do so if we set $p(C_1) = p(C_2) = 1/2$ and $\sigma_1 = 1$ $\sigma_2 = 1$?

Answer:

① Start with the assumption

$$p(x|C_1) = \mathcal{N}(x; \mu_1, \sigma_1^2)$$

and $$p(x|C_2) = \mathcal{N}(x; \mu_2, \sigma_2^2)$$

From the previous answer we know that given exponential distributions.

$$p(x|C_1) = h(x) \exp\left[\phi^T(x)\, \omega_1 - \ln Z(\omega_1)\right]$$

$$p(x|C_2) = h(x) \exp\left[\phi^T(x)\, \omega - \ln Z(\omega_2)\right]$$

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

as $a(x) = \ln\left[\frac{p(C_1|x)}{p(C_2|x)}\right]$

$$a(x) = \phi^T(x)\underbrace{(\omega_1 - \omega_2)}_{\theta} \underbrace{- \ln\frac{Z(\omega_1)}{Z(\omega_2)} + \ln\frac{p(C_1)}{p(C_2)}}_{\theta_0}$$

we know that the normal distribution also belongs to the exponential family. (Lecture-5, slide 13)

$$\mathcal{N}(x; \mu, \sigma^2) = \exp\left(\left[x \quad -\frac{x^2}{2}\right]\begin{bmatrix}\mu/\sigma^2 \\ 1/\sigma^2\end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi\sigma^2}\right)\right)$$

where $\phi^T(x) = [\phi_1(x) \quad \phi_2(x)] = \left[x \quad -\frac{x^2}{2}\right]$

$$\omega = \begin{bmatrix}\omega_1 \\ \omega_2\end{bmatrix} = \begin{bmatrix}\mu/\sigma^2 \\ 1/\sigma^2\end{bmatrix}$$

$$\ln Z(\omega) = -\frac{1}{2}\left[\frac{\omega_1}{\omega_2} - \ln\omega_2 + \ln(2\pi)\right]$$

$$= -\left[\frac{\mu^2}{2\sigma^2} + \ln(\sqrt{2\pi\sigma^2})\right]$$

Combining previous two equations we get

$$\frac{p(C_1|x)}{p(C_2|x)} = \exp\left\{ \begin{bmatrix} x & -\frac{x^2}{2} \end{bmatrix} \begin{bmatrix} \mu_1/\sigma_1^2 - \mu_2/\sigma_1^2 \\ 1/\sigma_1^2 - 1/\sigma_2^2 \end{bmatrix} - \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} - \ln \frac{(2\pi\sigma_1^2)}{(2\pi\sigma_2^2)} \right) + \ln \frac{p(C_1)}{p(C_2)} \right\}$$

Let us assume $p(C_1) = p(C_2) = 1/2$ and $\sigma_1 = \sigma_2 = \sigma$

then

$$\frac{p(C_1|x)}{p(C_2|x)} = \exp\left\{ \begin{bmatrix} x & -\frac{x^2}{2} \end{bmatrix} \begin{bmatrix} \mu_1/\sigma^2 - \mu_2/\sigma^2 \\ 0 \end{bmatrix} - \frac{1}{2}\left(\frac{\mu_1^2}{\sigma^2} - \frac{\mu_2^2}{\sigma^2} \right) \right\}$$

$$a(x) = \frac{\mu_1 - \mu_2}{\sigma^2} x - \frac{\mu_1^2 - \mu_2^2}{2\sigma^2}$$

$$a(x) = \left(\frac{\mu_1 - \mu_2}{\sigma^2} \right) \left[x - \frac{(\mu_1 + \mu_2)}{2} \right]$$

Clearly $\mu_1 = 3$, $\mu_2 = 0$, $\sigma^2 = 3$
will give $a(x) = x - \frac{3}{2}$

and $\mu_1 = 2$, $\mu_2 = 1$, $\sigma^2 = 1$
will result also result in $a(x) = x - 3/2$

Thus there does not exist a unique inverse mapping for an parameters θ, θ_0 to $\mu_1, \mu_2, \sigma_1, \sigma_2$

② Now consider the case $p(C_1) = p(C_2) = 1/2$

$\sigma_1 = \sigma_2 = \sigma = 1$

$$\frac{p(C_1|x)}{p(C_2|x)} = \exp\left\{ \begin{bmatrix} x & -\frac{x^2}{2} \end{bmatrix} \begin{bmatrix} \mu_1 - \mu_2 \\ 0 \end{bmatrix} - \frac{1}{2}\left(\mu_1^2 - \mu_2^2\right) \right\}$$

$$a(x) = (\mu_1 - \mu_2)\, x - \frac{1}{2}\left(\mu_1^2 - \mu_2^2\right)$$

Here given $\theta_1 = \mu_1 - \mu_2$

$$\theta_0 = -\frac{1}{2}\mu_1^2 - \mu_2^2 = -\frac{1}{2}\theta_1(\mu_1 + \mu_2)$$

or $\mu_1 - \mu_2 = \theta_1$

$\mu_1 + \mu_2 = -2\theta_0/\theta_1$

$$2\mu_1 = \theta_1 - \frac{2\theta_0}{\theta_1} = \frac{\theta_1^2 - 2\theta_0}{\theta_1}$$

$$\mu_1 = \frac{\theta_1^2 - 2\theta_0}{2\theta_1} \qquad \mu_2 = \frac{\theta_1^2 - 2\theta_0}{2\theta_1} - \theta_1 = -\frac{2\theta_0 - \theta_1^2}{2\theta_1}$$

Thus assuming $\theta_1 \neq 0$ μ_1 & μ_2 are recoverable from $a(x)$ given θ_1 and θ_0 from the result of logistic regression.