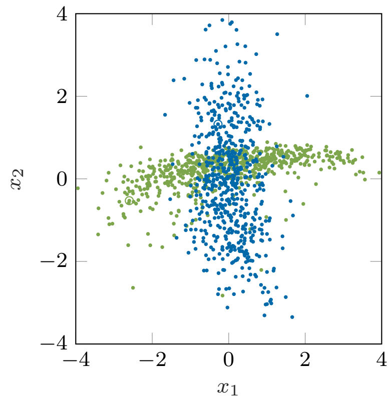
Exercise Sheet #9

due on Monday, 3 July 2023, 10am sharp

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 situtations indicate an *anti-causal* relationship ("the label causes the inputs"), and it might seem that the discriminative modeling paradigm adopted in the lecture is not appropriate. The goal of this week's theory exercise is to realise that the situation is a bit more subtle than that. For the purposes of this exercise, we will assume that there are two classes $\mathcal{C}_1, \mathcal{C}_2$ defining probability distributions $p(\boldsymbol{x} \mid \mathcal{C}_1), p(\boldsymbol{x} \mid \mathcal{C}_2)$ over the *inputs*, and classes are drawn with probability $p(\mathcal{C}) = [p(\mathcal{C}_1), p(\mathcal{C}_2) = 1 - p(\mathcal{C}_1)]$



   (a) Given a new input $\boldsymbol{x}$, how would you compute the posterior $p(\mathcal{C}_1 \mid \boldsymbol{x})$? Show that it can be written as a logistic function

$$p(\mathcal{C}_1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{-a(\boldsymbol{x})}} \quad \text{where} \quad a(\boldsymbol{x}) = \ln \frac{p(C_1 \mid \boldsymbol{x})}{p(C_2 \mid \boldsymbol{x})} \text{ are the } \textit{log odds}.$$

   (b) This observation suggests that when we perform logistic regression to learn the function $a(\boldsymbol{x})$, we actually indirectly learn the class distributions $p(\boldsymbol{x} \mid \mathcal{C}_1), p(\boldsymbol{x} \mid \mathcal{C}_2)$. It's interesting to consider how different regression models for $a(\boldsymbol{x})$ relate to different assumptions about the class distributions: Assume that both classes are a draws from the same exponential family, with different parameters:

$$p(\boldsymbol{x} \mid \mathcal{C}_k) = h(\boldsymbol{x}) \exp(\phi(\boldsymbol{x})^\intercal \boldsymbol{w}_k - \log Z(\boldsymbol{w}_k))$$

   Show that this implies a *linear* model for $a(\boldsymbol{x})$:

$$a(\boldsymbol{x}) = \phi(\boldsymbol{x})^\intercal \boldsymbol{\theta} + \theta_0$$

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

   (c) Does this mean we can turn any discriminative model (predicting classlabels from inputs, $p(\mathcal{C}_k \mid \boldsymbol{x})$) into a generative one (predicting inputs from class labels, $p(\boldsymbol{x} \mid \mathcal{C}_k)$) by "going backward" through the model? Unfortunately, the answer is *no*, because continuous distributions for inputs are more complex than binary distributions. To make this clear, consider a special case: Assume that the class distributions are both univariate Gaussians with different means and variances: $p(\boldsymbol{x} \mid \mathcal{C}_k) = \mathcal{N}(\boldsymbol{x}; \mu_k, \sigma_k^2)$. From your answer above, construct the explicit form of θ, θ_0 (that's two real numbers) as a function of $\mu_1, \mu_2, \sigma_1, \sigma_2, p(\mathcal{C}_1)$. Assume someone has performed logistic regression and given you the parameters θ, θ_0. Can you recover the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, p(\mathcal{C}_1)$? Could you do so if we set $p(C_1) = p(C_2) = \frac{1}{2}$ and $\sigma = 1$?

2. **Practical Question:** can be found in `Ex09.ipynb`