Probabilistic Generative models

Week 1 – Foundations

A PGM, or simply a GM, is a methodology for generating data. In general, the GM is constructed and adjusted using observations with the aim to generate data with the same statistical properties of the observations. The 'probabilistic' nature of the PGMs considered follow from the fact that the data generated will be considered to be realisations of a random variable, e.g. X.

In this sense, the probability distribution of X, denoted $P_X(x)$, as well as its density $p_X(x)$ will be central to the study of PGM. In particular, targeting the pdf is one way of constructing PGMs, in which case the whole PGM paradigm becomes equivalent to the classical statistical model approch. However, as we will see in the course, enforcing the sought-after PGM to have an explicit parametric pdf can be rather restrictive.

Throughout the course, we will consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with 3 RVs given by the measurable maps

$$\underset{\text{(observed input)}}{X:\Omega\to\mathcal{X}}\;;\;\underset{\text{(observed output)}}{Y:\Omega\to\mathcal{Y}}\;;\;\underset{\text{(latent variable)}}{Z:\Omega\to\mathcal{Z}}$$

Remark: Not all three variables will be present in all our models. For instance, in classification there's no justification for Z, while in clustering there's no need for Y. However, we build the general set up here for formality.

We will also consider the σ-algebra $\mathcal{F}$ to be the product Borel σ-algebre. $\mathcal{F} = \mathcal{B}(X) \otimes \mathcal{B}(Y) \otimes \mathcal{B}(Z)$.

Furthermore, we will assume the joint probability of X, Y and Z has a density, that is, $\forall A \in \mathcal{B}(X)$, $B \in \mathcal{B}(Y)$, $C \in \mathcal{B}(Z)$, we have:

$$\mathbb{P}(X \in A, Y \in B, Z \in C) = \int_{A \times B \times C} p(x,y,z)\,dx\,dy\,dz$$

We will also assume that all marginals and conditionals have a density. This includes $p(x,y)$, $p(y|x)$, $p(z|x,y)$ etc.

Discriminative v/s Generative. The generative approach aims to characterise the complete generative distribution $p(x,y,z)$, whereas in some application specific cases only the discriminative model $p(y|x)$ is needed. Let's se the following example.

Example: generative v/s discriminative binary classification.

Consider the bi classification problem, where given an observation $X=x$, one need's to estimate its label Y. A discriminative model would directly construct $p(Y|X=x)$. Since this is a binary case, without loss of generality we can assume $Y \in \{0,1\}$, and model $\mathbb{P}(Y=1|X=x)$, since $\mathbb{P}(Y=0|X=x) = 1-\%$ A model for this will only need to map $x \in \mathbb{R}^n \to \mathbb{P}(Y=1|X=x) \in [0,1]$. For instance

$$\mathbb{P}(Y=1|X=x) = \frac{1}{1+e^{-\theta^T x}}$$

which is known as the logistic regression.

Conversely, in a generative approach, we aim to model the joint probability $p(Y=1, X=x)$.

Modelling this distribution directly is not easy, however, we can observe that it can be factorised as

$$p(Y=1, X=x) = p(X=x \mid Y=1)\, p(Y=1)$$

which yields two much more intuitive distributions:

- the class prob $p(Y=1) \to \pi, (1-\pi)$
- the class-conditional prob $p(X=x / Y=1) \to$ choose… $f_{\theta_1}, f_{\theta_2}$

Therefore, the classifier is

$$p(Y=1 \mid X=x) = \frac{p(X=x \mid Y=1)\, p(Y=1)}{p(X=x)}$$

$$= \frac{1}{1 + \dfrac{p(X=x \mid Y=0)\, p(Y=0)}{p(X=x / Y=1)\, p(Y=1)}}$$

$$= \frac{1}{1 + e^{\log\left(\frac{1-\pi}{\pi} \cdot \frac{f_{\theta_0}(x)}{f_{\theta_1}(x)}\right)}} \qquad \circledast$$

Exercise: evaluate $\circledast$ for $f_{\theta_1} = N(\mu_1, \Sigma)$, $f_{\theta_0} = N(\mu_2, \Sigma)$

- MLE example
- push forwards
- inf theo
- pdf not always possible