Machine Learning Assignment 10

Question 1

Consider a forward SDE:

$$dx_t = f(x_t, t)dt + g(x_t, t)dW_t,$$

show that the corresponding probability flow ODE is written as

$$dx_t = \left[f(x_t, t) - \frac{1}{2}\frac{\partial}{\partial x} g^2(x_t, t) - \frac{g^2(x_t, t)}{2}\frac{\partial}{\partial x}\log p(x_t, t) \right] dt$$

(Pf.)

In forward SDE: $f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ and $g(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$.

We have already know that the corresponding Fokker-Planck equation is

$$\frac{\partial p}{\partial t} = -\nabla \cdot (fp) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2}{\partial x_i \partial x_j}[g(x_t, t)g(x_t, t)^T p]$$

If we write it in element-wise, that is

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}[f_i(x_t, t)p(x)] + \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left[\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\left(\sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)p(x)\right)\right]$$

Note that the latter term

$$\begin{aligned}
&\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\left(\sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)p(x)\right) \\
&\overset{\text{(Product rule)}}{=} \sum_{j=1}^{d}\left\{\frac{\partial}{\partial x_j}\left[\sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)\right]p(x) + \sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)\cdot\frac{\partial}{\partial x_j}(p(x))\right\} \\
&\overset{\frac{\partial}{\partial x_j}\ln p(x) = \frac{1}{p}\frac{\partial p}{\partial x_j})}{=} \sum_{j=1}^{d}\left\{\frac{\partial}{\partial x_j}\left[\sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)\right]p(x) + \sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)\cdot p(x)\frac{\partial}{\partial x_j}(\ln p(\right. \\
&= \sum_{j=1}^{d}\frac{\partial}{\partial x_j}\left[\sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)\right]p(x) + \sum_{j=1}^{d}\sum_{k=1}^{d} g_{ik}(x_t, t)g_{jk}(x_t, t)\cdot p(x)\frac{\partial}{\partial x_j}(\ln p(x)) \\
&= p(x)\nabla\cdot(g(x_t, t)g(x_t, t)^T) + p(x)g(x_t, t)g(x_t, t)^T\cdot\nabla(\ln p(x))
\end{aligned}$$

Continue writing, we have

$$\begin{aligned}\frac{\partial p}{\partial t} &= -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}[f_i(x_t,t)p(x)] \\ &\quad + \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left[p(x)\nabla\cdot(g(x_t,t)g(x_t,t)^T) + p(x)g(x_t,t)g(x_t,t)^T\cdot\nabla(\ln p(x))\right] \\ &= -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left\{f_i(x_t,t)p(x) - \frac{1}{2}\left[p(x)\nabla\cdot(g(x_t,t)g(x_t,t)^T) + p(x)g(x_t,t)g(x_t,t)^T\cdot\nabla(\ln p(x\right.\right. \\ &= -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left\{f_i(x_t,t)p(x) - \frac{1}{2}\left[\nabla\cdot(g(x_t,t)g(x_t,t)^T) + g(x_t,t)g(x_t,t)^T\cdot\nabla(\ln p(x))\right]p(x\right. \\ &:= -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\tilde{f}_i(x_t,t\end{aligned}$$

where we define

$$\tilde{f}_i(x_t,t) = f_i(x_t,t) - \frac{1}{2}\nabla\cdot(g(x_t,t)g(x_t,t)^T) - \frac{1}{2}g(x_t,t)g(x_t,t)^T\cdot\nabla(\ln p(x))$$

That is, $\frac{\partial p}{\partial t} = \nabla\cdot(\tilde{f}p)$, which is a SDE without diffusion term. By Liouville equation, its corresponding ODE is $dx_t = \tilde{f}dt$. And in $d=1$ case, that is the desired result

$$dx_t = \left[f(x_t,t) - \frac{1}{2}\frac{\partial}{\partial x}(g^2(x_t,t)) - \frac{g^2(x_t,t)}{2}\frac{\partial}{\partial x}\ln p(x,t)\right]dt$$

Question 2

AI 對於目前的考古學領域已經發揮了一定的影響力，例如能處理大量的地形或衛星資料、辨識古文物的材質以推測年代、利用自然語言處理破譯古文字等等。但同時也有一些瓶頸急待解決，例如：

- 同一件文物在不同的背景下可能有不同的解釋（像是同時具有宗教、技術、貿易用途），無法簡單的只用一類標籤去標記。
- 承上，文物上可能留下一些符號與文字，AI 可能可以幫助我們辨認文字，但無法給予解釋。且像是古符號通常帶有主觀解釋空間，對於 AI 來說要理解意義似乎還有難度。
- 雖然透過影像辨識技術能認出物體，但因為考古資料通常是多模態的，包含影像、地層、年代測定、出土紀錄、文本記載等等。因此，以現在最多只能處理一到兩種模態的 AI 模型來說，難以整合時間、空間與文化資訊。
- 最後，由於決策過程很常是黑箱，所以 AI 給出的結果很常是無法解釋的。考古學家不一定能接受沒有原因的推論結果。
- 或者，若 AI 給出錯誤的判斷結果，有可能造成文物盜掘的風險。

本次我想專注在 「未來 20 年，AI 除了能辨認古器上面的文字或紋理之外，還能幫助專家分類並標註，並從全部的未標註樣本中挑選最值得專家親自標註的樣本」。
這其中的瓶頸包括：

- 資料大部分破碎且不完整（像是古器或是損壞的碑文）

- 仰賴深厚的背景知識，才能認出是哪個時期的文物
- 因為不同學者有不同詮釋，所以不一定有標準答案

若要達成這個目標，會需要以下的機器學習方法：

- Self-supervised learning (自監督式學習)
  - 讓模型能從未標註的資料中自我學習影像或文字特徵，不需依賴人工標記
- Unsupervised learning（無監督式學習）
  - 將出土器物的紋理或形式分群，以利後續命名或研究
- Few-shot learning（少樣本學習）
  - 因為古器數量不可能有非常多，若模型能用相對少數的樣本學習到特徵，那麼將較為省時且符合效益
- Active learning（主動學習）
  - 讓模型自動挑出最值得專家標註的部分，以此降低人工標註的比例

以上除了主動學習會由人類進行外部回饋，其餘三者的回饋皆只有內部回饋，也就是透過定義內部損失函數來優化整個學習過程。

那麼，我們現在設計一個簡化過的模型問題，以達到讓 AI 能幫助學者標註並分類古文物的目標。我們先只以古陶片為對象，希望完成以下任務：

1. 圖樣/形狀分類：輸入陶片照片 x^{image} 與其邊緣/輪廓 x^{shape}，並預測其形狀及紋理類別 y^{cls}。預期類別會很多而每個類別內的資料數量很少
2. 紋理分割：對同一張照片的每個像素輸出類別 $y_{ij} \in \{0, 1, \cdots, K\}$，不同類別代表不同區域（背景、器物本體、兩種以上紋理、缺損等等）
3. 時期/文化歸類：結合地點座標 x^{geo}、地層順序 x^{strat} 及文字紀錄片段 x^{text} 來預測時期 y^{period}
4. 新型態偵測：若判斷屬於目前未擁有的未知類，則特別標記 $y^{unknown}$ 提示專家進行標記
5. 不確定度：對每個預測產生可信度，並依此從未標註樣本中挑選最值得讓專家標註的樣本

若這些任務能達成，某種程度上也代表了實際的能力。因為真實考古時樣本數應不多，且若全部要藉由人工標註所費不貲，並且同時要分辨器型、區域標註及推測年代時期。且將影像、文字、地理與地層座標等資料都餵入模型，讓模型去學多模態的資訊整合。新型態偵測也模擬了真實會發生的情況。

針對這個小實驗可以設計一些指標來評斷模型是否成功：

- Top-1 Accuracy：模型預測的第一名類別對應正確類別的比例
- Top-5 Accuracy：模型預測的前五名類別包含正確類別的比例
- AUROC：衡量模型能否區分「已知類別」與「未知類別」的能力，愈接近 1 愈好
- mIoU：預測區域與真實標註區域的交集面積 / 聯集面積，衡量紋理分類的準度
- 區間重疊率：計算模型預測年代區間與實際區間的重疊率
- ECE：分成多個可信度區間，檢查預測可信度與實際正確率差距的平均

- Brier Score：測量模型預測結果的整體信心與真實結果的接近程度
- 違反地層/年代/地理座標限制的比例：用來衡量模型預測是否遵守考古邏輯的指標

其中，可能涉及到的機器學習技術與數學如下：

1. 對比學習：讓模型學會「若兩片陶片有相似的特徵則距離近，不同的則要分開」，令函數

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\text{sim}(z_i, z_j^+)/\tau)}{\sum_k \exp(\text{sim}(z_i, z_k)/\tau)}$$

   其中 $\text{sim}(\cdot,\cdot)$ 為餘弦相似度，若兩個特徵向量相似則夾角小，輸出值愈接近 1，τ 控制分布平滑度

2. Few-shot learning 少樣本學習：讓模型在極少樣本下辨識新器型類別。建立每類的中心

$$c_k = \frac{1}{|S_k|} \sum_{\{x_i : y_i = k\}} f_\theta(x_i)$$

   ，則對於新樣本 x，以距離決定類別：

$$P(y = k|x) = \frac{\exp(-\|f_\theta(x) - c_k\|^2)}{\sum_j \exp(-\|f_\theta(x) - c_j\|^2)}$$

3. Neuro-symbolic Integration：讓模型推論遵守考古邏輯，例如地層 A 比地層 B還早，就可以設計 Loss 為：

$$\mathcal{L}_{rule} = \max(0, \mu_A - \mu_B + \delta)$$

   ，當預測的年代違反地層順序關係（$\mu_A - \mu_B > 0$）就會被懲罰

4. Uncertainty Sampling：在所有未標註樣本中，找出最不確定的那些樣本請專家標註，因為這些樣本能提供最多新資訊，才值得人工標註。可以透過定義熵的方式：

$$x^* = \text{argmax}_x(-\sum_k P(y = k|x) \log P(y = k|x))$$

   來找到最值得讓專家人工標註的那些樣本

Question 3

> Unanswered questions

When performing linear interpolation between two latent noise samples $x_T^{(1)}$ and $x_T^{(2)}$, and then running the backward ODE (or probability flow ODE) to generate an image, should we always expect the resulting image to visually resemble a "mixture" of the two source images?

For instance, if $x_T^{(1)}$ and $x_T^{(2)}$ correspond to a dog image and a cat image respectively, is it possible that the generated interpolated image does not resemble either a dog or a cat?

(Texts are refined by ChatGPT.)