

20240327 人工智慧與商業應用

w6

彙納方法 (Inductive methods) 是指一種科學或研究方法，其特點包括：

通過觀察和分析來推斷一般性結論的科學方法，其適用於那些無法從已知原理中直接得出結論的情況。

觀察與資料收集：以觀察現象、事件或資料為起點，並收集相關的數據。

從特例到一般性：從具體的實例或觀察開始，推斷出一般性的規律或模式。

基於結果建立理論：通過收集大量資料、觀察行為、建立模型或提出假設，然後根據這些資料和觀察的結果來制定普遍性的理論或推斷。

非嚴格的推理過程：與演繹方法相比，彙納方法的推理過程通常更不確定，因為它基於概率和可能性，而不是嚴格的邏輯推理。

演繹法：

優點：

1. **邏輯嚴謹：**演繹法基於嚴格的邏輯推理，因此結論通常具有高度的確定性。
2. **快速推斷：**使用現有的一般原理進行推斷，因此能夠迅速地得出結論。

缺點：

1. **依賴於前提：**演繹法的結論僅取決於已知的前提，如果前提不正確或不完整，則可能導致錯誤的結論。
2. **不適用於未知情況：**演繹法難以應用於解決新問題或探索未知領域，因為它僅限於現有的知識和原理。

歸納法：

優點：

1. **適用於未知情況**：歸納法能夠從特定觀察中推斷**一般性規律**，因此適**用於解決新問題或探索未知領域**。
2. **彈性**：歸納法的推理過程較為彈性，**能夠應對複雜和不確定的情況**。

缺點：

1. **確定性較低**：歸納法的結論通常具有較低的確定性，因為它**受限於樣本數據的代表性和觀察的主觀性**。
2. **容易受到誤導**：**受到樣本選擇偏差或觀察偏見的影響**，從而導致**不準確的結論**。

演繹法是從一般原理推斷特定結論的推理方法，具有嚴謹的邏輯和高度確定性。舉例來說，如果我們知道「所有人會死亡」（一般原理），並且「小明是人」（特殊前提），那麼我們可以演繹出結論：「小明將會死亡」。

歸納法則是從特定觀察推斷一般性規律的推理方法，具有彈性但確定性較低。例如，觀察到過去幾個夏季都有大量的降雨，可以歸納出「這個地區夏季降雨量較高」的結論。

Vector Space Model (VSM) 向量空間模型

用於訊息檢索與文本分析，會用到TF-IDF的概念(那個權重表格)

以找工作為例：

1. 建立詞彙空間

- 收集與就業相關的詞彙，包括職位名稱、技能、工作地點、行業、公司名稱等。
- 將這些詞彙組成詞彙集合，作為向量空間的基底

2. 文檔表示

- 將每份工作的工作描述或職位要求轉換為向量表示。
- 對於每個工作，計算每個詞彙在工作描述中的詞頻或TF-IDF權重，作為向量的相應維度值。

3. 查詢表示

- 將您的職業技能、經驗、求職意願等轉換為向量表示。

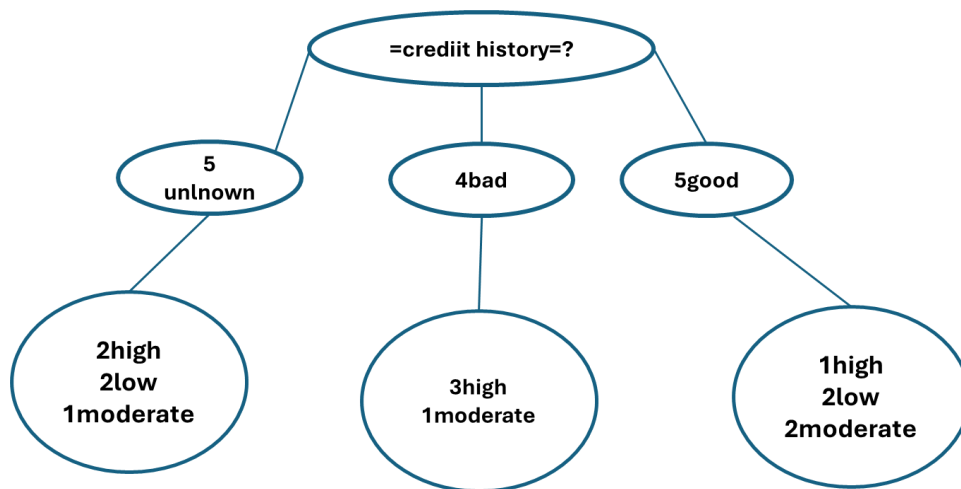
- 對於您的查詢，計算每個詞彙在您的表示中的詞頻或TF-IDF權重。

4. 相似度計算

- 使用相似度計算方法（如餘弦相似度），計算每份工作描述向量與您的查詢向量之間的相似度。
- 根據相似度排名，找到與您的求職意願最相似的工作。

就可能會建立像這個表一樣的資料庫(或資料集dataset)

| | Y | X ₁ | X ₂ | X ₃ | X ₄ |
|-----|----------|----------------|----------------|----------------|----------------|
| | RISK | Credit history | Debt | Collateral | Income |
| E1 | High | Bad | High | None | <15k |
| E2 | High | Unknown | High | None | 15k-35k |
| E3 | Moderate | Unknown | Low | None | 15k-35k |
| E4 | High | Unknown | Low | None | <15k |
| E5 | Low | Unknown | Low | None | >35k |
| E6 | Low | Unknown | Low | Adequate | >35k |
| E7 | High | Bad | Low | None | <15k |
| E8 | Moderate | Bad | Low | Adequate | >35k |
| E9 | Low | Good | Low | None | >35k |
| E10 | Low | Good | High | Adequate | >35k |
| E11 | High | Good | High | None | <15k |
| E12 | Moderate | Good | High | None | 15k-35k |
| E13 | Low | Good | High | None | >35k |
| E14 | High | Bad | High | None | 15k-35k |



Entropy 熵

degree of **chaos** for a distribution 是分布的混亂程度

熵可以用來衡量一個消息的不確定性或意外性。

例如，**如果一個事件是必然發生的，那麼它的熵就是零；但如果一個事件是完全隨機的，那麼它的熵就會很高**

熵是一個物理學和信息理論中的概念，用來描述一個系統的混亂程度或不確定性。在物理學中，**熵是指一個系統內部的無序程度，通常越高表示系統越混亂。**

我們可以依照公式去求出平均的entropy 並想盡辦法降低entropy 這樣才能使的資料正確的判斷產出正確

回到剛剛的表格，原本使用credit history不能帶來最小的entropy

我們要決定誰當根