

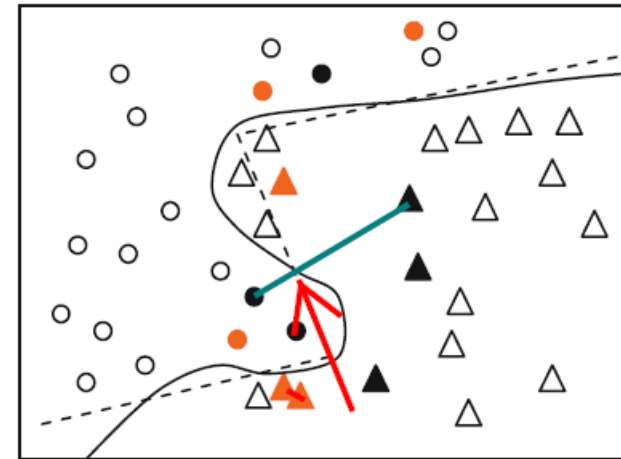
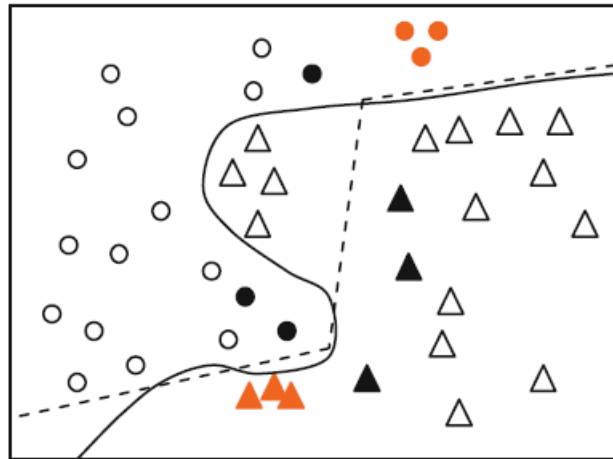
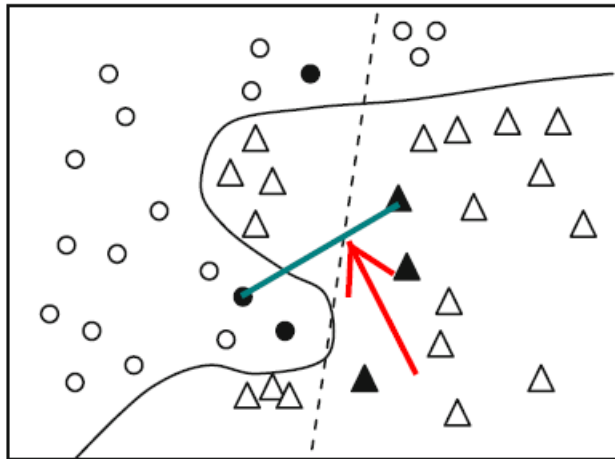
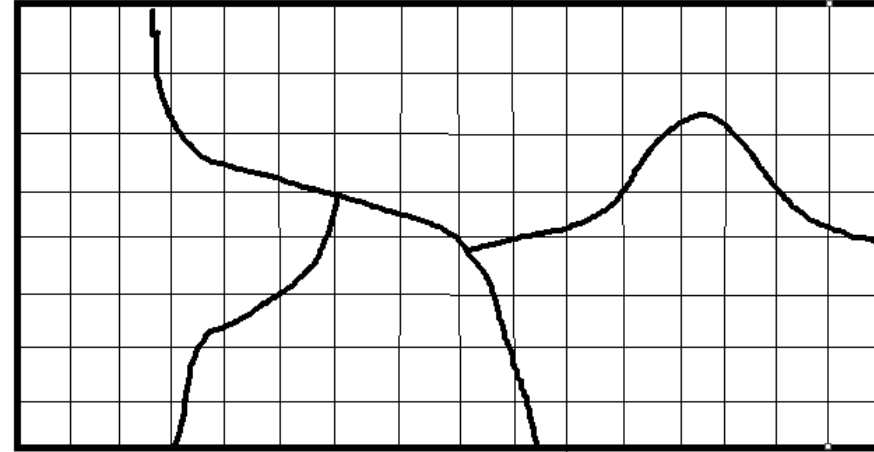


Large Scale Active Learning with Generative Model about Uncertainty Data

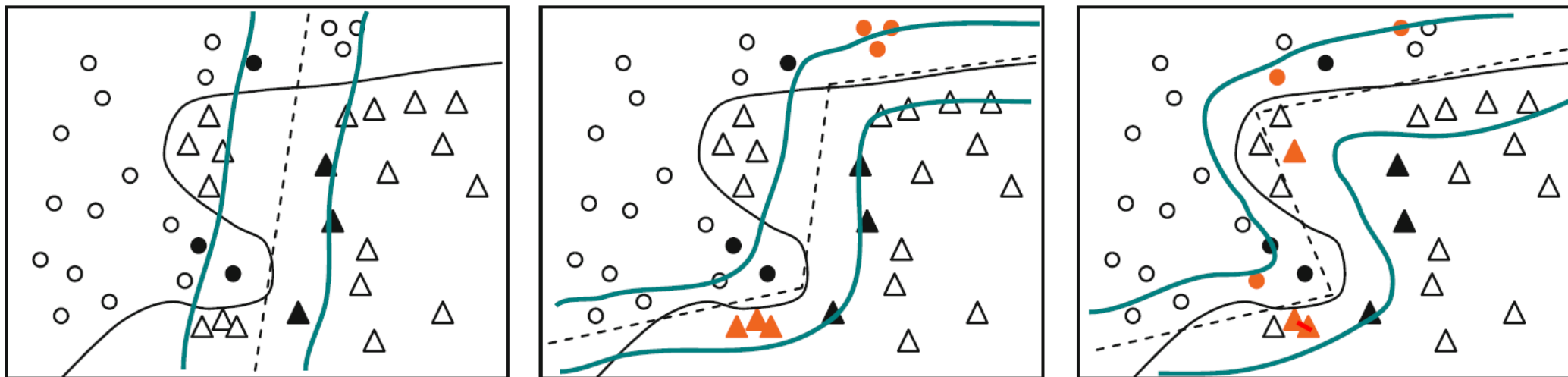
2017.10.20

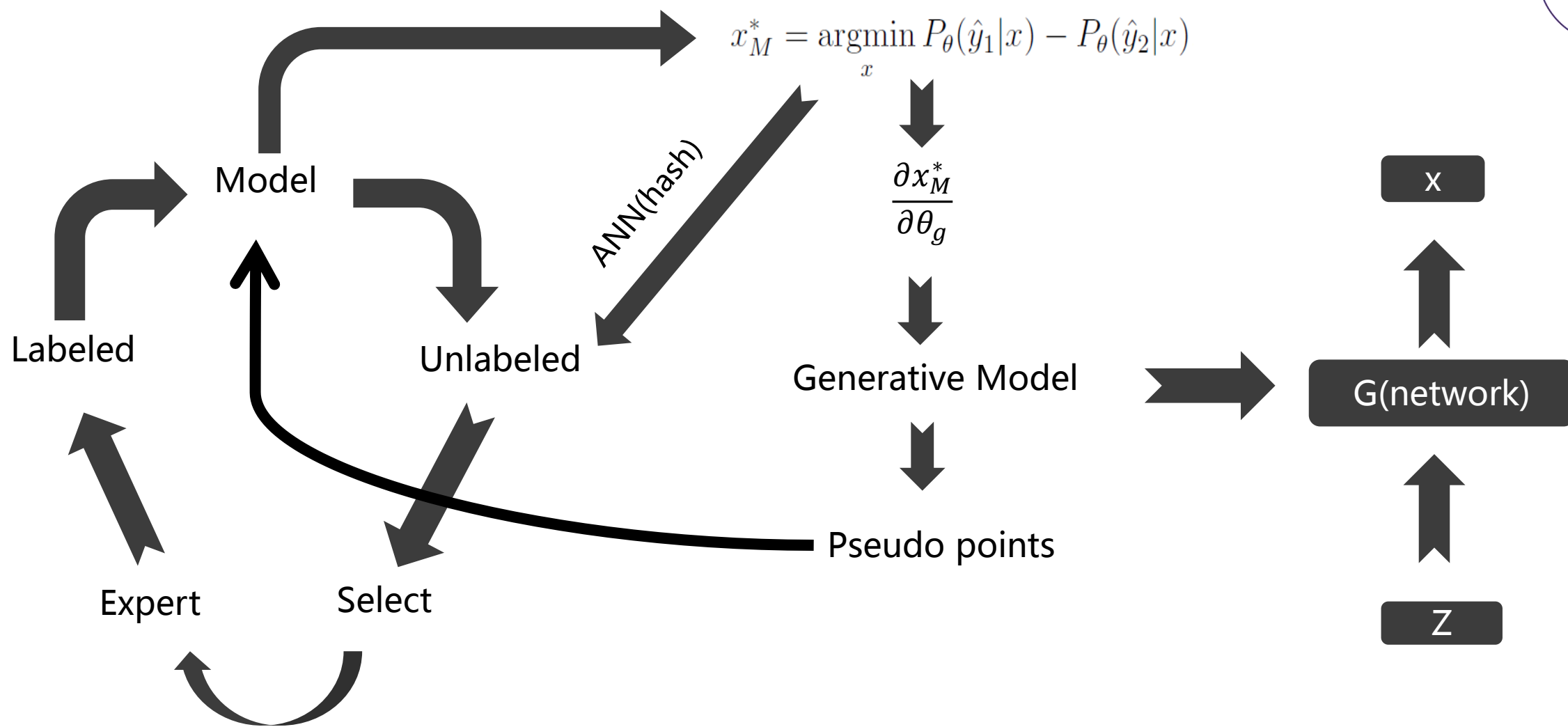
1.如何显式或隐式表示（控制）
分类界面

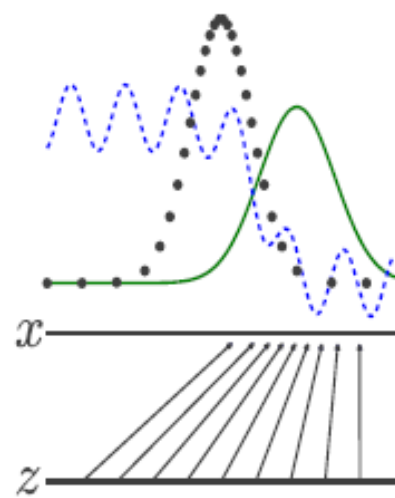
2.如何描述（或知晓）新查询一个点后分类
界面的变化



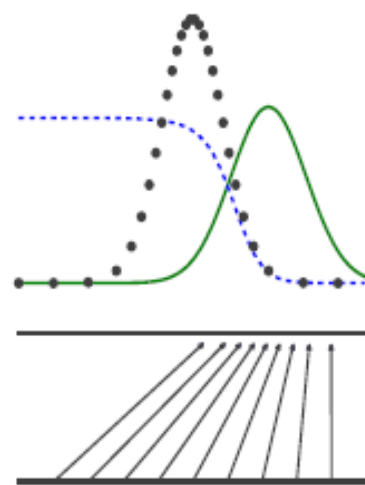
描述当前分类器分类边界附近的带状区域



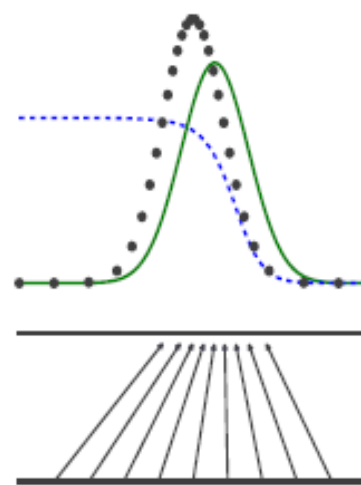




(a)

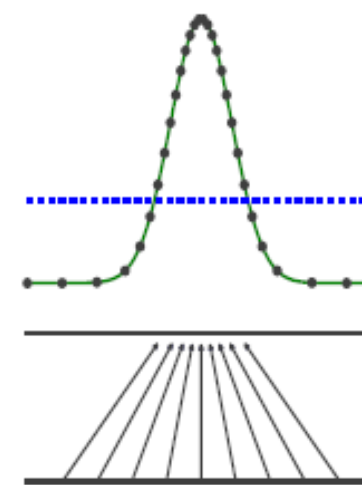


(b)



(c)

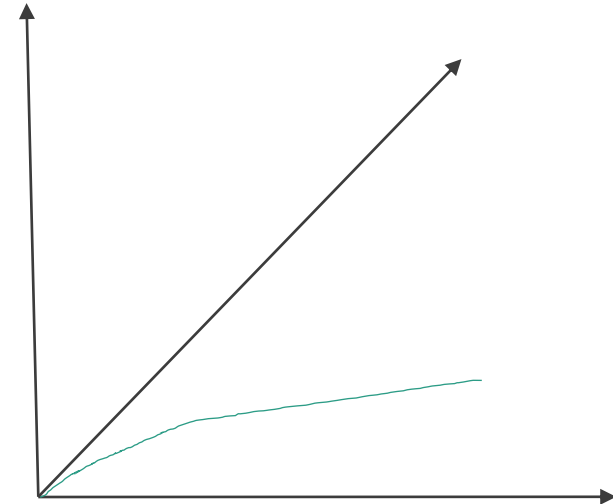
...



(d)

Advantages

- 1 在query过t个数据后，生成模型G的参数 $\theta_{g|t}$ 作为下次迭代中生成模型参数的初始值，由于在Active Learning训练中后期分类边界变化不大，所以该初始值已接近最优，节省了G的训练时间。（即训练成本越来越小）
- 2 当Label数量确定，随数据集规模线性增加，生成模型训练成本缓慢增加。
- 3 分类边界变化可描述性。
- 4 AL中分类模型通用性，以及生成模型通用性。



Problem

- ❶ 目标函数选择及反向传播求导运算。 (Uncertainty Sampling)

$$x_{LC}^* = \operatorname{argmax}_x 1 - P_{\theta}(\hat{y}|x) \quad x_H^* = \operatorname{argmax}_x - \sum_i P_{\theta}(y_i|x) \log P_{\theta}(y_i|x) \quad x_M^* = \operatorname{argmin}_x P_{\theta}(\hat{y}_1|x) - P_{\theta}(\hat{y}_2|x)$$

- ❷ 验证生成模型基本可行性。 (二分类logistic)



- ❸ 更换常用实际大规模数据集，更换模型，多分类，是否在大规模实际数据集上有效。