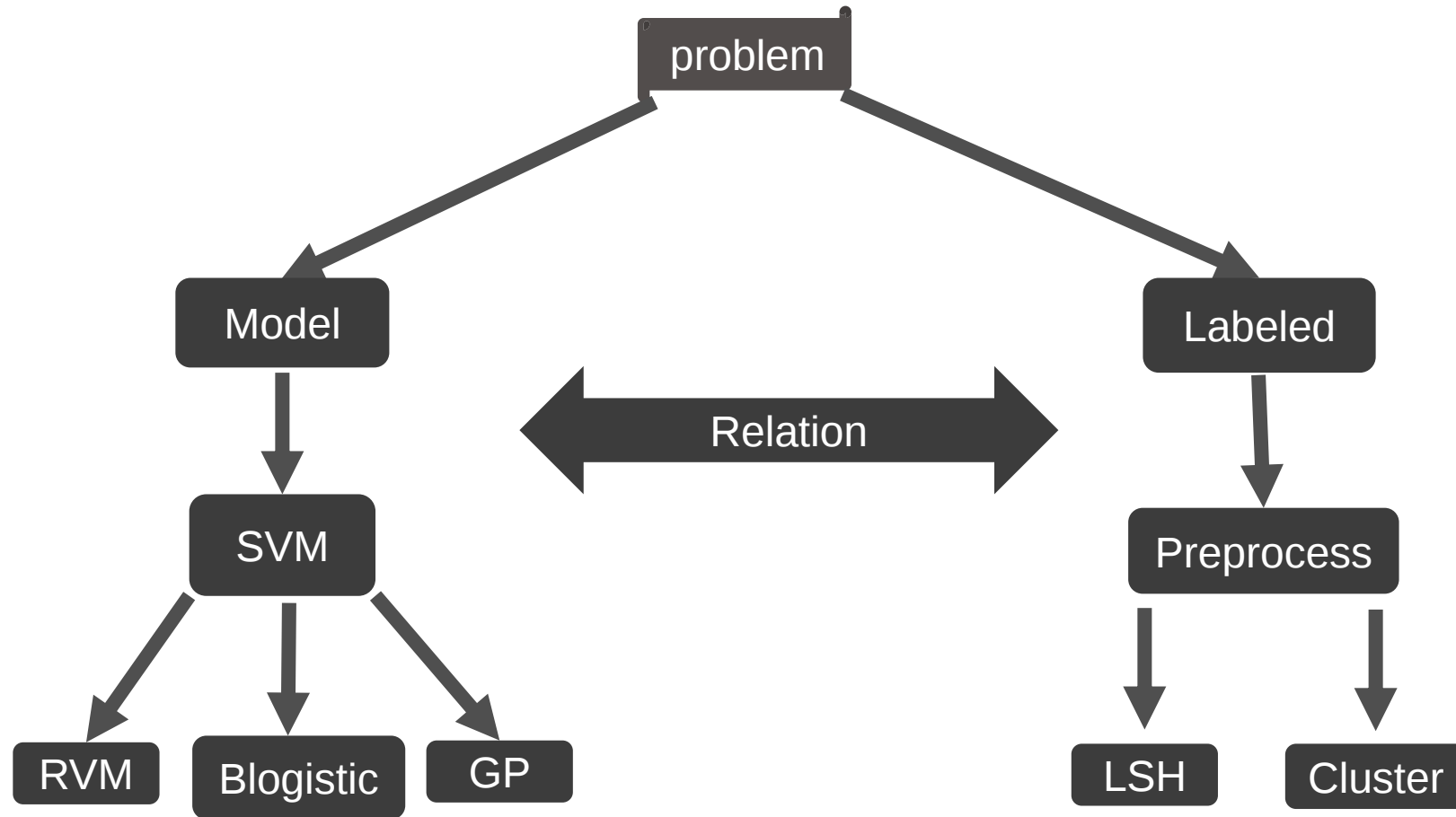




Discussion About Large Scale Active Learning

2017.9.25





- 1.如何显式或隐式表示（控制）分类界面
- 2.如何描述新查询一个点后分类界面的变化
- 3.如何从已标记数据点推断分类边界位置继而寻找下一查询点位置
- 4.如何判断模型部分边界已经最优（AL趋于收敛）
- 5.数据维度和编码问题



1.如何隐式表示（控制）分类界面

从模型角度根据所建模型依据寻找边界 $p(y_i|\theta, x) = 0.5$

$$p(C_1 | \phi, \mathbf{t}) = \sigma(\kappa(\sigma_a^2)\mu_a) \quad \kappa(\sigma^2) = (1 + \frac{\pi\sigma^2}{8})^{-\frac{1}{2}} \quad \sigma_a^2 = \phi^T \mathbf{S}_N \phi \quad \mu_a = \mathbf{w}_{MAP}^T \phi$$

2.如何描述新查询一个点后分类界面的（局部）变化

查询一个点后，模型参数的变化量：新模型相对于上次模型对所有已标记数据点的预测确信度变化

或直接考虑参数关于数据点及标记的关系： $\theta_{i+1} = f(x_{i+1}, y_i)$

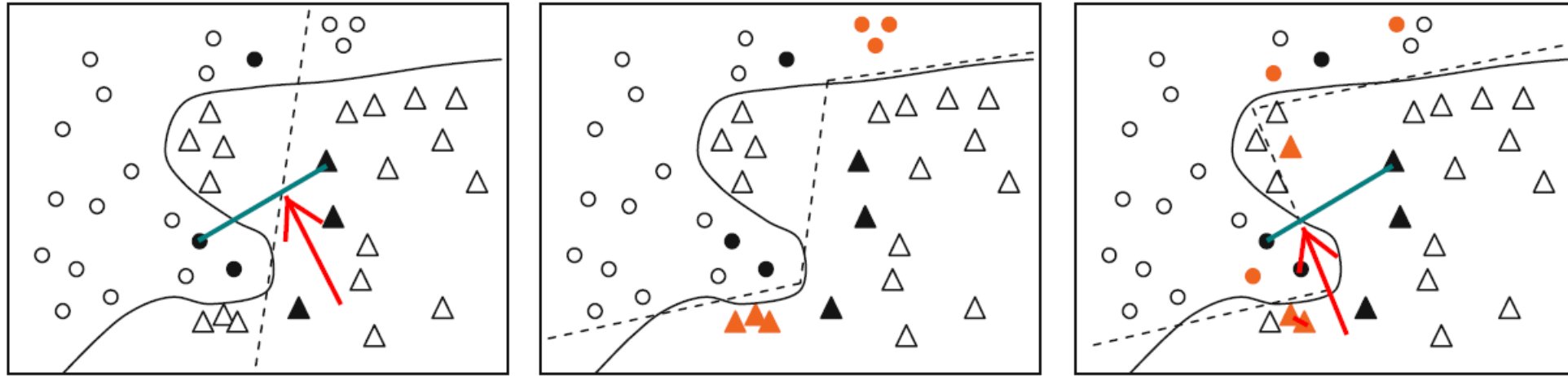


引入LSH (Locality-Sensitive Hashing) ,根据伪数据点锁定真实数据点

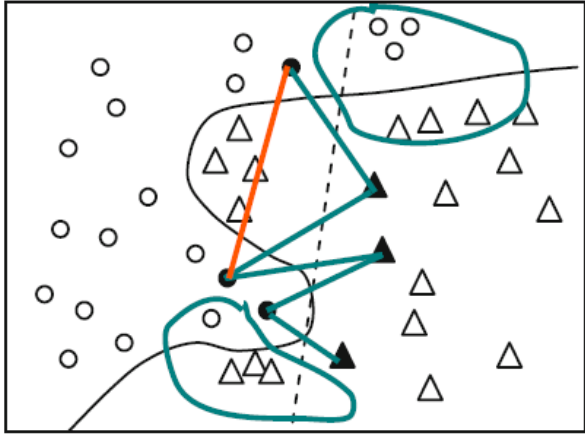
$$x = F^{-1}(p = 0.5, \theta) \longrightarrow x^{data}$$

3.如何从已标记数据点推断分类边界位置继而寻找下一查询点位置

4.如何判断模型部分边界已经最优（如何判断AL收敛）

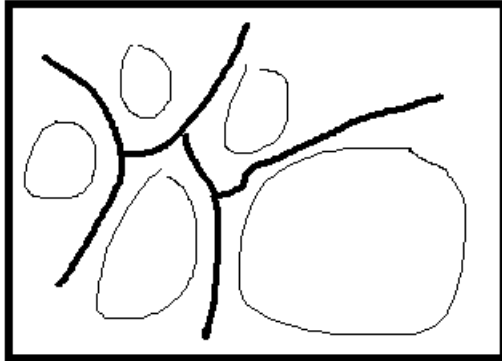


例如：二分查找不确定度最大位置即分类界面（注：获得的位置通常为伪数据点）



回归模型建立+数据预处理：考虑聚类结构或ANN（LSH）

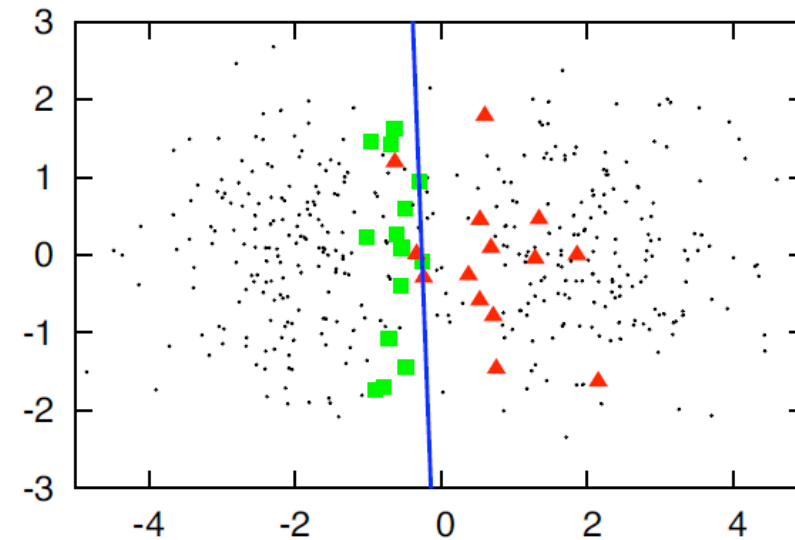
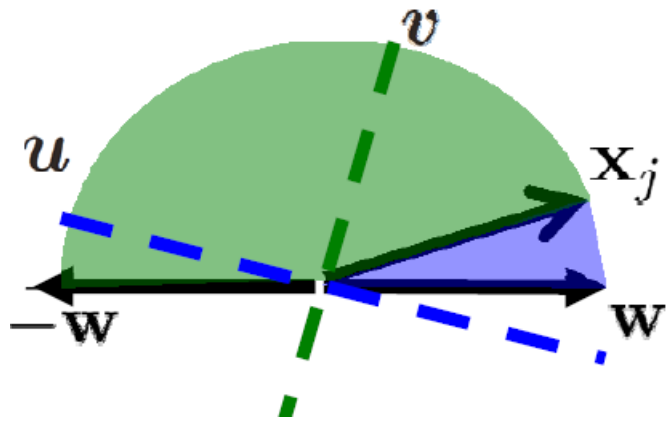
根据已标记数据点对分类边界建模，根据模型在不同位置的不确定度大小和离聚类中心的距离定位查询点位置。



回归模型在不同标记的数据点之间穿插，根据回归模型推出回归线上的点并根据LSH确定出距离聚类中心较近的点

5. 数据维度和编码问题

Without loss of generality, we assume that the hyperplane passes through origin, and that each x_i , ω is unit norm.





Accelerated Large Scale Optimization by Concomitant Hashing



Min-Product

Max-Product

$$\mathcal{X} = \left\{ x \mid |\omega^T x| \leq |\omega^T x'|, \forall x' \notin \mathcal{X} \right\} \quad \mathcal{X} = \left\{ x \mid |\omega^T x| \geq |\omega^T x'|, \forall x' \notin \mathcal{X} \right\}$$

$$\min_{y \in \{-1,1\}^N, b, \xi \geq 0} \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \quad s.t. \quad y_i (w^T x_i + b) \geq 1 - \xi_i, \forall i,$$

$$(\text{Lasso}) : \min_{\alpha} \|x - D\alpha\|^2 + \lambda \|\alpha\|_1.$$

$$(\text{OMP}) : \min_{\alpha} \|x - D\alpha\|^2, \quad s.t. \quad \|\alpha\|_0 \leq k.$$

$$(\text{GPR}) : \min_{\alpha} \frac{1}{2} \alpha^T (K + \sigma^2 I) \alpha - y^T \alpha.$$