Positive-only feedback

解释:

- 更关心正类样本，符合 PU 数据的生成过程。
- 负类种类多样
- 标注要求相对较低，不易犯错:
    - 一方面，只有当很确定的情况下才将样本标注为正类，将负类标注为正类的可能性相对较小 (Noise);
    - 另一方面，即使漏掉正类也不会造成太大的影响 (PU)。

主动学习：增加标注数据。在 PU Learning 中，标注数据只能是正类，因此，如何能增加最有用的正类数据是 Active PU Learning 要考虑的。

Problem

- Two samples 问题:
  - 从 U 中标注数据违背了这一假设
  - P 和 U 的分布在标注过程中会改变，正类先验（正类在 U 中的占比）相应地也会改变

假定：U 集合足够大，每次查询中少量正类从 U 移至 P 后对 U 的分布影响不大
在迭代过程中，根据查询到的正类数目修改 prior 的值

为什么prior是必须的?
- Learning from positive and unlabeled data.
- 只知道什么是正类，但是没有被告知什么是非正类
- 学习 binary classifier 需要额外的条件：prior

Select Queries

PU Learning 中哪些样本对分类器准确率影响较大

决定性因素是数目较少的 P 集合。标注什么样的正类样本能够最大程度地提升分类器性能?

直观上而言:

1) 最具代表性/多样性的正类
2) 不确定度最大的正类

ERM

$$\arg\min_{f\in H_k} \frac{\pi}{n_P}\sum_{x_i\in X_P} l(f(x_i),+1)-l(f(x_i),-1)+\frac{1}{n_U}\sum_{x_j\in X_U} l(f(x_j),-1)$$

$$l(f(x_i),+1)-l(f(x_i),-1)=f(x_i)$$

$$=\arg\min_{f\in H_k} \frac{\pi}{n_P}\sum_{x_i\in X_P} f(x_i)+\frac{1}{n_U}\sum_{x_i\in X_U} l(f(x_i),-1)$$

$$=\arg\min_{f\in H_k} \widehat{R}_P+\widehat{R}_U$$

$$g(f,P,U)=\frac{\pi}{n_P}\sum_{x_i\in X_P} f(x_i)+\frac{1}{n_U}\sum_{x_i\in X_U} l(f(x_i),-1)+\frac{\lambda}{2}\|f\|_{H_k}^2$$

$$\arg\min_{S\in U \wedge |S|=b}\min_{f\in H_k} \widehat{R}_{P\cup S_+}+\widehat{R}_{U\setminus S_+}$$

$$\arg\min_{q^T 1_U=b}\min_{f\in H_k} \tilde{g}(f,P,U,q)$$

q: 被选中的概率

p: 被当前分类器预测为正类的概率

$$\tilde{g}(f,P,U,q)=\frac{\pi}{n_P}\sum_{x_i\in X_P} f(x_i)+\frac{1}{n_U}\sum_{x_j\in X_U} q_j p_j \pi f(x_i)+(1-q_j)(1-p_j)l(f(x_j),-1)+\frac{\lambda}{2}\|f\|_{H_k}^2$$

$$\arg\min_{q^T 1=b} \min_{f\in H_k} \tilde{g}\left(f,P,U,q\right)$$

$$\begin{aligned}
\tilde{g}(f,P,U,q) &= \frac{\pi}{p}\sum_{x_i\in X_P} f(x_i) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_i) + (1-q_j)(1-p_j) l(f(x_j),-1) + \frac{\lambda}{2}\|f\|_{H_k}^2 \\
&= \frac{\pi}{p}\sum_{x_i\in X_P} f(x_i) + \frac{1}{u}\sum_{x_j\in X_U} l(f(x_j),-1) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_j) + (q_j p_j - q_j - p_j) l(f(x_j),-1) + \frac{\lambda}{2}\|f\|_{H_k}^2 \\
&= R\left(f\right) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_i) + (q_j p_j - q_j - p_j) l(f(x_j),-1) + \frac{\lambda}{2}\|f\|_{H_k}^2 \\
&\le g(f^{'},P,U) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_i) + (q_j p_j - q_j - p_j) l(f(x_j),-1) \\
&= g(f^{'},P,U) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_i) - q_j(1-p_j) l(f(x_j),-1) - \frac{1}{u}\sum_{x_j\in X_U} p_j l(f(x_j),-1)
\end{aligned}$$

向量化

$$f(x_j)=\sum_{i:x_i\in D}\alpha_i k(x_i,x_j)$$

$$\frac{1}{n_U}\sum_{x_j\in X_U} q_j p_j \pi f(x_i)-q_j(1-p_j)\frac{(f(x_j)+1)^2}{4}-\frac{1}{n_U}\sum_{x_j\in X_U} p_j l(f(x_j),-1)+\frac{\lambda}{2}\|f\|_{H_k}^2$$

$$=\frac{1}{n_U}\sum_{x_j\in X_U} q_j p_j \pi\left(\sum_{x_i\in D}\alpha_i k(x_i,x_j)\right)-\frac{q_j(1-p_j)}{4}\left\{\left(\sum_{x_i\in D}\alpha_i k(x_i,x_j)\right)^2+2\left(\sum_{x_i\in D}\alpha_i k(x_i,x_j)\right)+1\right\}$$

$$-\frac{1}{n_U}\sum_{x_j\in X_U} p_j\left\{\left(\sum_{x_i\in D}\alpha_i k(x_i,x_j)\right)^2+2\left(\sum_{x_i\in D}\alpha_i k(x_i,x_j)\right)+1\right\}$$

$$=\frac{\pi}{n_U}\langle q\circ p\rangle^T K_{U,D}\alpha-\frac{1}{4n_U}\alpha^T K_{U,D}^T\langle q\circ(1-p)\rangle\langle q\circ(1-p)\rangle^T K_{U,D}\alpha-\frac{1}{2n_U}\langle q\circ(1-p)\rangle^T K_{U,D}\alpha-\frac{1}{4n_U}\langle q\circ(1-p)\rangle$$

$$-\frac{1}{n_U}\alpha^T K_{U,D}^T p p^T K_{U,D}\alpha-\frac{2}{n_U}p^T K_{U,D}\alpha-\frac{1}{n_U}p$$

$$\frac{\pi}{n_U}\langle q\circ p\rangle^T K_{U,D}\alpha-\frac{1}{4n_U}\alpha^T K_{U,D}^T\langle q\circ(1-p)\rangle\langle q\circ(1-p)\rangle^T K_{U,D}\alpha-\frac{1}{2n_U}\langle q\circ(1-p)\rangle^T K_{U,D}\alpha-\frac{1}{4n_U}\langle q\circ(1-p)\rangle$$

$$\arg\min_{q^T 1=b} \min_{f\in H_k} \tilde{g}\left(f, P, U, q\right)$$

$$\begin{aligned}
\tilde{g}(f, P, U, q) &= \frac{\pi}{p}\sum_{x_i\in X_P} f(x_i) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_i) + (1-q_j)(1-p_j) l(f(x_j), -1) + \frac{\lambda}{2}\|f\|_{H_k}^2 \\
&= \frac{\pi}{p}\sum_{x_i\in X_P} f(x_i) + \frac{1}{u}\sum_{x_j\in X_U} l(f(x_j), -1) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_j) + (q_j p_j - q_j - p_j) l(f(x_j), -1) + \frac{\lambda}{2}\|f\|_{H_k}^2 \\
&= R\left(f\right) + \frac{1}{u}\sum_{x_j\in X_U} q_j p_j \pi f(x_i) + (q_j p_j - q_j - p_j) l(f(x_j), -1) + \frac{\lambda}{2}\|f\|_{H_k}^2 \\
&\le g(f', P, U) + \frac{1}{u}\sum_{x_j\in X_U} p_j\left(q_j \pi f(x_i) + (q_j - 1) l(f(x_j), -1)\right) - q_j l(f(x_j), -1) \\
&= g(f', P, U) + \frac{1}{u}\sum_{x_j\in X_U} p_j\left(q_j \pi\left(l(f(x_j), +1) - l(f(x_j), -1)\right) + q_j l(f(x_j), -1) - l(f(x_j), -1)\right) - \frac{1}{u}\sum_{x_j\in X_U} q_j l(f(x_j), -1) \\
&= g(f', P, U) + \frac{1}{u}\sum_{x_j\in X_U} p_j\left(q_j\left(R_{x_j}^{+} + R_{x_j}^{-}\right) - l(f(x_j), -1)\right) - \frac{1}{u}\sum_{x_j\in X_U} q_j l(f(x_j), -1)
\end{aligned}$$

Heuristic

$$\arg\min_{\alpha^T 1_U = b} \boxed{MMD(P \cup B, \hat{P})} + \boxed{\frac{dist(P, B)}{\Sigma_{P \cup B}}}$$

Trade-off

希望 batch 中的正类样本多一些而且具有多样性