1 1

1.1 Bayesian Decision Theory[Lin et al. 2002]

Training data가 joint distibution $P(\mathbf{X}, Y)$에서 뽑힌 i.i.d. sample이라 가정합니다. Bayeisan decision theory는 이 $P(\mathbf{X}, Y)$를 알고 있을 때 최적의 classifier를 찾아내는 기준입니다. 이 최적의 classifier를 Bayes' rule이라고 합니다. Bayes' rule을 정의하기 위해 먼저 아래의 값들을 정의합니다.

[Risk] c^+ = cost of false negative, c^- = cost of false positive
[Prior] $\pi^+ = Pr(Y = 1)$, $\pi^- = Pr(Y = -1)$
[Likelihood] $g^+(\mathbf{x}) = Pr(\mathbf{X} = \mathbf{x}|Y = 1)$, $g^-(\mathbf{x}) = Pr(\mathbf{X} = \mathbf{x}|Y = -1)$
[Posterior] $p(\mathbf{x}) = Pr(Y = 1|\mathbf{X} = \mathbf{x}) = \dfrac{\pi^+ g^+(\mathbf{x})}{\pi^+ g^+(\mathbf{x}) + \pi^- g^-(\mathbf{x})}$
classifier f는 아래와 같을 때 Bayes' rule입니다.

$$\text{for given } \mathbf{x},\ f(\mathbf{x}) = \begin{cases} = 1, \text{ if } c^+ p(\mathbf{x}) > c^-(1 - p(\mathbf{x})) \quad i.e. \quad \dfrac{p(\mathbf{x})}{1 - p(\mathbf{x})} > \dfrac{c^-}{c^+} \\ -1, \text{ otherswise.} \end{cases}$$

1.2 Resampling[Lin et al. 2002, Liu et al. 2010]

Ideal resampling은 c^+, c^- ratio를 바꾸는 것과 같은 효과를 낸다는 것을 보이겠습니다. 여기서 ideal resampling은 likelihood를 보존하면서 prior만 변경하는 것을 의미합니다. 예를 들어, 질병 관련 연구를 위해 환자와 비환자 집단에서 샘플 100명을 뽑는다고 할 때, 전체 인구 집단에서 환자의 비율이 0.1, 비환자 비율이 0.9라면 prior는 $\pi^+ = 0.1, \pi^- = 0.9$입니다. 의도적으로 환자 집단에서 50명, 비환자 집단에서 50명을 random sampling하면 likelihood는 그대로이면서 prior만 $\pi^+ = \pi^- = 0.5$로 변경되고 이에 따라 posterior도 변경됩니다.

Ideal resampling을 거친 sample은 더 이상 $P(\mathbf{X}, Y)$의 i.i.d. sample이라 볼 수 없습니다. Resampled sample의 population distribution을 $P_s(\mathbf{X}, Y)$라 하고, 이에 해당하는 prior를 π_s^+, π_s^-라 하겠습니다. 이를 통해 새로운 posterior $p_s(\mathbf{x})$를 아래와 같이 계산합니다.

$$p_s(\mathbf{x}) = \frac{\pi_s^+ g^+(\mathbf{x})}{\pi_s^+ g^+(\mathbf{x}) + \pi_s^- g^-(\mathbf{x})}$$

이에 기반한 Bayes' rule은

$$\text{for given } \mathbf{x},\ f(\mathbf{x}) = \begin{cases} = 1, \text{ if } \dfrac{p_s(\mathbf{x})}{1 - p_s(\mathbf{x})} > \dfrac{c^-}{c^+} \\ -1, \text{ otherswise.} \end{cases}$$

입니다. 즉, resampled sample 기반으로 학습한 최적 classifier는 원본 sample 로 학습한 최적 classifier와 다릅니다. 하지만, 둘 사이에는 아래와 같은 간단한 관계가 성립합니다.

$$\begin{aligned}\frac{p_s(\mathbf{x})}{1-p_s(\mathbf{x})} > \frac{c^-}{c^+} &\iff \frac{\pi_s^+ g^+(\mathbf{x})}{\pi_s^- g^-(\mathbf{x})} > \frac{c^-}{c^+} \\ &\iff \frac{\pi^+ g^+(\mathbf{x})}{\pi^- g^-(\mathbf{x})} \cdot \frac{\pi_s^+ \pi^-}{\pi_s^- \pi^+} > \frac{c^-}{c^+} \\ &\iff \frac{\pi^+ g^+(\mathbf{x})}{\pi^- g^-(\mathbf{x})} > \frac{c^-}{c^+}\frac{\pi_s^- \pi^+}{\pi_s^+ \pi^-}\end{aligned}$$

즉, resampled data로 학습한 Bayes' rule은 원본 데이터에서 $\frac{c^-}{c^+}$를 $\frac{\pi_s^- \pi^+}{\pi_s^+ \pi^-}$ 만큼 수정하여 학습한 Bayes' rule과 같습니다.

Class imbalance를 개선하기 위해 resampling을 사용했을 때 위의 결과를 해석해 보면 다음과 같습니다. Positive sample이 적은 상황을 가정합니다. 원래의 imbalance ratio를 $i = \frac{\pi^-}{\pi^+}$로 정의하고, resampling 이후 개선된 imbalance ratio를 $i_s = \frac{\pi_s^-}{\pi_s^+}$로 정의하면, $\frac{\pi_s^- \pi^+}{\pi_s^+ \pi^-} = \frac{i_s}{i}$, 즉 imbalance ratio가 개선된 정도를 의미합니다. 예를 들어, $i = 10$에서 resampling을 통해 $i_s = 5$로 개선했다면 $i_s = \frac{1}{2}$가 $\frac{c^-}{c^+}$에 곱해지는 것입니다. 이는 결국 c^+를 2배 올리는 것과 같습니다.

1.3 Asymptotically fixing the effect of resampling using WSVM [Lin 1999, Lin et al. 2002]

Lin 1999에서 SVM의 asymptotic target이 Bayes' rule임을 보였습니다. Lin et al. 2002에서 resampling이 Bayes' rule에 미치는 영향을 WSVM을 통해 상쇄할 수 있음을 보였습니다. SVM의 objective function에서 slack variable ξ_i에 아래의 weight $L(y_i)$를 곱하면, resample한 데이터로 학습한 SVM의 asymptotic target은 원본 데이터 기반 Bayes' rule과 같습니다.

$$\begin{cases} L(1) = c^+ \pi^+ \pi_s^- \\ L(-1) = c^- \pi^- \pi_s^+ \end{cases}$$

1.4 Resampling for highly imbalanced big data

1. SVM은 big data에 취약하다. kernel SVM은 sample size n에 대해 시간 복잡도가 $O(n^3 * p)$이다. → Intel 4850HQ processor (Core i7 2.3

GHz quadcore)에서 n=1000000, p = 2 일 경우 약 40초 소요. High-dimensional일 경우 상당히 많은 시간이 소요될 것으로 예상. 게다가 svm은 paramter tuning이 '필수'이므로, 여러 C와 gamma 조합을 시도해 봐야 하기 때문에 training time이 짧아야 한다. 만약 10개의 gamma 값과 10개의 C값을 시도한다면, training time은 100배 증가합니다. Prediction time도 decision boundary의 시각화 등에 필요하기 때문에 짧은 것이 좋다.

2. 시간을 줄이는 방법은 undersampling.

3. imbalance를 해결하는 방법은 WSVM, undersampling, oversampling.