Supplementary material to the paper Convex Multiple-Instance Learning by Estimating Likelihood Ratio

October 5, 2010

1 Proof of Theorem 1

Lemma 1 *If $2x$ assumes a beta distribution with parameter $(1, \beta)$, the expection of x, x^2, x^3, x^4 is*

$$\begin{aligned}
\mathbb{E}(x) &= \frac{1}{2}\frac{1}{\beta+1} \\
\mathbb{E}(x^2) &= \frac{1}{2}\frac{1}{(\beta+2)(\beta+1)} \\
\mathbb{E}(x^3) &= \frac{3}{4}\frac{1}{(\beta+1)(\beta+2)(\beta+3)} \\
\mathbb{E}(x^4) &= \frac{3}{2}\frac{1}{(\beta+1)(\beta+2)(\beta+3)(\beta+4)}
\end{aligned}$$

Proof: It is known that the expectation of the beta distribution $B(1, \beta)$ is $\mathbb{E}(x) = \frac{1}{\beta+1}$, therefore the first equation is immediate.

For $k > 1$, the expectations $\mathbb{E}(x^k)$ of a beta distribution can be computed as:

$$\begin{aligned}
& \int_0^1 \frac{1}{B(\alpha,\beta)} x^k x^{\alpha-1}(1-x)^{\beta-1} dx \\
= & \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^k x^{\alpha-1}(1-x)^{\beta-1} dx \\
= & \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k-1)}{\Gamma(\alpha+k-1+\beta)\Gamma(\alpha)} \int_0^1 \frac{\Gamma(\alpha+k-1+\beta)}{\Gamma(\alpha+k-1)\Gamma(\beta)} x^k x^{\alpha-1}(1-x)^{\beta-1} dx \\
= & \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k-1)}{\Gamma(\alpha+k-1+\beta)\Gamma(\alpha)} \int_0^1 x \frac{1}{B(\alpha+k-1,\beta)} x^{\alpha+k-2}(1-x)^{\beta-1} dx \\
= & \frac{\prod_{j=0}^{k-2}(\alpha+j)}{\prod_{j=0}^{k-2}(\alpha+\beta+j)} \frac{\alpha+k-1}{\alpha+k-1+\beta} (k \geq 2) \qquad (1)
\end{aligned}$$

The last step use the formulae $B(x+y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ and $\Gamma(x+1) = x\Gamma(x)$, as well as the expectation of the beta distribution $\mathbb{E}(x) = \frac{\alpha}{\alpha+\beta}$.

If $2x$ instead of x assumes the beta distribution, then

$$\begin{aligned}\mathbb{E}(x^k) &= \frac{1}{2^k}\mathbb{E}((2x)^k) \\ &= \frac{1}{2^k}\frac{\prod_{j=0}^{k-2}(\alpha+j)}{\prod_{j=0}^{k-2}(\alpha+\beta+j)}\frac{\alpha+k-1}{\alpha+k+\beta-1}. \end{aligned} \quad (2)$$

Substite $\alpha = 1, k = 2, 3, 4$ we obtain the Lemma. □

Proof of Theorem 1:

Proof: Define $\eta(x_i) = \Pr(y = 1|x_i, \Pr(y = -1|x_i) > \Pr(y = 1|x_i))$. The most adversarial scenario is $\Pr(1 - 2\eta(x_i) \leq 2\epsilon) = c\epsilon^\beta$, because this maximizes uniformly the chance that the class-conditional probability is close to $1/2$. This means $2\eta(x_i)$ assumes a beta distribution $B(1, \beta)$. It could be proved that the likelihood ratio satisfies

$$\frac{9}{4}\eta^2 + \frac{3}{4}\eta \leq \frac{\eta}{1-\eta} \leq 2\eta^2 + \eta$$

From Lemma 1, we know that $E(\eta^2) = \frac{1}{2}\frac{1}{(\beta+2)(\beta+1)}, E(\eta) = \frac{1}{2}\frac{1}{\beta+1}$, therefore we have

$$\frac{3}{8}\frac{\beta+5}{(\beta+1)(\beta+2)} \leq \mathbb{E}(\frac{\eta}{1-\eta}) \leq \frac{\beta+4}{2(\beta+1)(\beta+2)}$$

Furthermore, a bound on the variance can be computed as

$$\begin{aligned} V(\frac{\eta}{1-\eta}) &= \mathbb{E}[(\frac{\eta}{1-\eta})^2] - \mathbb{E}[(\frac{\eta}{1-\eta})]^2 \\ &\leq 4E(\eta^4) + 4E(\eta^3) + E(\eta^2) - \frac{9}{64}\frac{(\beta+5)^2}{(\beta+1)^2(\beta+2)^2} \\ &\leq (\frac{1}{2} + \frac{5}{8(\beta+1)} + \frac{51\beta-83}{8(\beta+1)(\beta+3)(\beta+4)})\frac{1}{(\beta+1)(\beta+2)} \end{aligned}$$

in which $E(\eta^4) = \frac{3}{2}\frac{1}{(\beta+1)(\beta+2)(\beta+3)(\beta+4)}$, $E(\eta^3) = \frac{3}{4}\frac{1}{(\beta+1)(\beta+2)(\beta+3)}$. The last step involves some quite complicated arithmetics and simplification.

A simpler bound can be expressed as:

$$V(\frac{\eta}{1-\eta}) \leq \frac{(4\beta+1)}{2(\beta+1)^2(2\beta+3)},$$

which is an upper bound of the above bound.

By Bennett's inequality [1],

$$\frac{1}{n}\sum_{i=1}^{n} Z_i - E(Z) \leq \sqrt{\frac{2V(Z)\log 1/\delta}{n}} + \frac{\log 1/\delta}{3n}$$

therefore with probability $1-\delta$,

$$\sum_{i=1}^{n} Z_i \leq nE(Z) + \sqrt{2nV(Z)\log 1/\delta} + \frac{\log 1/\delta}{3}.$$

Take $Z = \frac{\eta}{1-\eta}$ and plug in the mean and variance bounds computed previously, we obtain the theorem. □

2 Proof of Theorem 2

(a) It is straightforward to show that

$$\begin{aligned} R(f) - R^* &= R(f) - R(\eta - 1/2) \\ &= E(\mathbf{1}[f(X) > 1, \eta(X) < 1/2](1 - 2\eta(X)) \\ &\quad + E(\mathbf{1}[f(X) < 1, \eta(X) > 1/2](2\eta(X) - 1)) \\ &= (R_-(f) - R_-^*) + R_+(f) - R_+^* \end{aligned}$$

We apply Jensen's inequality to both summands and obtain

$$\begin{aligned} & \psi(-\mathbf{1}[\eta(X) < 1/2](R(f) - R^*))) + \psi(\mathbf{1}[\eta(X) \geq 1/2](R(f) - R^*)) \\ \leq\ & E(\psi(-\mathbf{1}[\eta(X) < 1/2, f(X) > 1](1 - 2\eta(X))) \\ & + \psi(\mathbf{1}[\eta(X) \geq 1/2, f(X) < 1](2\eta(X) - 1))) \\ =\ & E(\mathbf{1}[\mathrm{sign}(f(X) - 1) \neq \mathrm{sign}(\eta(X) - 1/2)]\psi((2\eta(X) - 1))) \\ \leq\ & E(\mathbf{1}[\mathrm{sign}(f(X) - 1) \neq \mathrm{sign}(\eta(X) - 1/2)]\tilde{\psi}((2\eta(X) - 1))) \\ =\ & E(\mathbf{1}[\mathrm{sign}(f(X) - 1) \neq \mathrm{sign}(\eta(X) - 1/2)](H^-(\eta(X)) - H(\eta(X)))) \\ =\ & E\left(\mathbf{1}[\mathrm{sign}(f(X) - 1) \neq \mathrm{sign}(\eta(X) - 1/2)]\left(\inf_{\alpha,(\alpha-1)(2\eta(X)-1)\leq 0} C(\alpha, \eta(X)) - H(\eta(X))\right)\right) \\ \leq\ & E(C(\alpha, \eta(X)) - H(\eta(X))) \\ =\ & R_C(f) - R_C^* \end{aligned}$$

(b) First note that since $\psi(0) = 0$ and ψ is continuous, $\theta_i \to 0$ implies $\psi(\theta_i) \to 0$. Thus we can replace condition (2) by

(2') For any sequence (θ_i) in $[-1, 1]$,

$$\psi(\theta_i) \to 0 \text{ implies } \theta_i \to 0$$

To see that (1) implies (2'), let C be classification-calibrated, and let (θ_i) be a sequence that does not converge to 0. Define $c = \limsup \theta_i > 0$, and pass to a subsequence with $\lim \theta_i = c$. Then by continuity $\lim \psi(\theta_i) = \psi(c)$, and $\psi(c) > 0$ by classification-calibration. Thus for the original seequence (θ_i), we see $\limsup \psi(\theta_i) > 0$, so we cannot have $\psi(\theta_i) \to 0$.

To see that (2') implies (3), suppose that $R_C(f_i) \to R_C^*$. By part (a) of the theorem $\psi^-(R_-(f_i) - R_-^*) + \psi(R_+(f_i) - R_+^*) \to 0$, since both ψ^- and ψ are convex and positive, (2') implies $R(f_i) \to R^*$.

Finally, to see (3) implies (1), suppose C is not classification-calibrated. By definition, we can choose $\eta \neq 1/2$ and a sequence $\alpha_1, \alpha_2, \ldots$ such that $\mathrm{sign}(\alpha_i(\eta - 1/2)) = -1$ but $c_\eta(\alpha_i) \to H(\eta)$. Fix x and choose the probability distribution P so that $P_X\{x\} = 1$ and $P(Y = 1|X = x) = \eta$. Define a sequence of functions f_i for which $f_i(x) = \alpha_i$. Then $\lim R(f_i) > R^*$, and this is true for any infinite subsequence. But $C_\eta(\alpha_i) \to H(\eta)$ implies $R_C(f_i) \to R_C^*$. The contradiction proves the final part of the theorem. □

3 Equivalent Constant Transforms

Proposition 1 *For loss functions that satisfy a scaling equality $L(k_1\hat{y}, k_1 y) = k_2 L(\hat{y}, y)$, solving (4) with $C = C_0$ and D_i for each bag is equivalent with solving (3) with $C = \frac{k_1^2}{k_2} C_0$ and $k_1 D_i$.*

Proof: A variable substitution of $z = k_1 y$ and $v = k_1 w$ in (3) would obtain the conclusion. □

The support vector regression (SVR) we use satisfies $L(k_1\hat{y}, k_1 y; \epsilon) = k_1 L(\hat{y}, y; \frac{\epsilon}{k_1})$. In this case, the uniform scaling on all the D_is is equivalent to subsequent changes in both C and ϵ.

4 Projection to the Bag Constraint

We use the following procedure to project y^+: Denote the sum of scores in a bag as $s_i = \sum_{x_j^+ \in B_i} y_j^+$. For each bag B_i, we check if $s_i < D_i$, if not, we simply set all the negative y_j^+ to 0. If $s_i < D_i$, we add $\frac{D_i - s_i}{|B_i|}$ to each y_j^+. This makes $s_i = D_i$. Then we set all the negative y_j^+ to 0. If this makes $s_i > D_i$, we subtract equal amount on all the positive y_j^+, to make $s_i = D_i$. If this created additional negative y_j^+, the alternating projection process goes on until both conditions: $s_i = D_i$ and $y^+ \geq 0$ are fulfilled.

5 Derivation of the SVM dual problem

First rewrite the optimization in the canonical form:

$$\begin{aligned} \min_{w,b,y^+} \quad & \tfrac{1}{2}\|w\|^2 + C(\textstyle\sum_{i=1}^{n_+ + n_-}(\xi_i + \hat{\xi}_i)) \\ \text{s.t.} \quad & (\langle w, \phi(x_i)\rangle + b) - y_i \leq \epsilon + \xi_i, \text{for each } x_i^+, x_i^- \\ & y_i - (\langle w, \phi(x_i)\rangle + b) \leq \epsilon + \hat{\xi}_i, \text{for each } x_i^+, x_i^- \\ & \textstyle\sum_{x_j^+ \in B_i} y_j^+ \geq D_i |B_i| \\ & y_j^+ \geq 0 \end{aligned}$$

where $y_i = 0$ for x_i^- and $K(x, y) = \langle \phi(x), \phi(y)\rangle$.

Setting α_i^-, α_i^+ as the Lagrange multipliers of the first and second set of constraints, and η_j the Lagrange multipliers of the bag constraints, we obtain the following KKT condition:

$$\begin{gathered} w = \textstyle\sum_i (\alpha_i^+ - \alpha_i^-)\phi(x_i) \\ 0 \leq \alpha_i^+, \alpha_i^- \leq C \\ \eta_j \leq \alpha_i^+ - \alpha_i^- \\ \textstyle\sum_i (\alpha_i^+ - \alpha_i^-) = 0 \\ \alpha_i^+, \alpha_i^-, \eta_j \geq 0 \end{gathered}$$

From the KKT conditions one could see that to make the equality constraint $\sum_{x_j^+ \in B_i} y_j^+ = D_i|B_i|$ hold, we need $\eta_j > 0$, essentially, each α_i^+ in the bag must be positive. This means that $y_i - (\langle w, \phi(x_i)\rangle + b) \geq \epsilon$ for all the items in the bag. Thus, to make the equality hold, all items in the positive bag need to be predicted smaller than their real value y_i. Another issue is, since $\eta \geq 0$ must hold, $\alpha_i^+ \geq \alpha_i^-$, since only one of α_i^+ and α_i^- is nonzero, this means $\alpha_i^- = 0$ for instances in positive bags. Therefore, the situation that $\alpha_i^+ = 0, \alpha_i^- > 0$ could never exist. Back to the original optimization problem, this means $(\langle w, \phi(x_i)\rangle + b) - y_i \leq \epsilon + \xi_i$ never holds in equality. Therefore, $(\langle w, \phi(x_i)\rangle + b) \leq y_i + \epsilon$.

The dual problem is very similar with the original SVM, with only minor differences introduced from η:

$$\begin{aligned} \min_{\alpha^+, \alpha^-, \eta} \quad & \tfrac{1}{2}(\alpha^+ - \alpha^-)^T K(\alpha^+ - \alpha^-) \\ & + \epsilon \textstyle\sum_{i=1}^n (\alpha_i^+ + \alpha_i^-) - \sum_{i=1}^k D_i|B_i|\eta_i \\ \text{s.t.} \quad & 0 \leq \alpha_i^+, \alpha_i^- \leq C \\ & \textstyle\sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) = 0 \\ & \eta_j \leq \alpha_i^+ - \alpha_i^-, x_i \in B_j \\ & \alpha_i^+, \alpha_i^-, \eta_j \geq 0 \end{aligned}$$

From the dual problem, we could see that η_j needs to be made larger to improve the result of the dual. Therefore, the algorithm would always prefer to choose $\eta_j > 0$. From our previous analysis this essentially means that the equality constraint $\sum_{x_j^+ \in B_i} y_j^+ = D_i|B_i|$ is desirable. Thus the algorithm would tend to make more vectors from positive bags as support vectors. And tend to make predictions smaller than the estimated value y_i.

It could be seen that the best solution of η_j is $\eta_j = \max(\min_{x_i \in B_j}(\alpha_i^+ - \alpha_i^-), 0)$, with this in mind, the problem can still be solved using an SMO-type active set approach of selecting two α_i for one iteration.

The SMO subproblem is thus:

$$\begin{aligned}
\min_{\alpha_i,\alpha_j} \quad & \tfrac{1}{2}\begin{bmatrix} s_i\alpha_i & s_j\alpha_j \end{bmatrix}\begin{bmatrix} Q_{ii} & Q_{ij} \\ Q_{ij} & Q_{jj} \end{bmatrix}\begin{bmatrix} s_i\alpha_i \\ s_j\alpha_j \end{bmatrix} + \epsilon(\alpha_i + \alpha_j) \\
& -D_{k_1}|B_{k_1}|\eta_{k_1} - D_{k_2}|B_{k_2}|\eta_{k_2} \\
\text{s.t.} \quad & 0 \le \alpha_i, \alpha_j \le C \\
& s_i\alpha_i + s_j\alpha_j \text{ doesn't change}
\end{aligned}$$

where s_i and s_j are the signs of α_i and α_j, respectively.

The SMO working set selection would be the same as in the original SVM, i.e., find the maximal violating pair, except that the gradient now needs to take η into consideration.

References

[1] Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58** (1963) 13–30