Optimistic Classification

Anti-Boosting or Robust Classification based on Phd Marina Agullo

E. del Barrio & J.-M. Loubes*

15 february 2018, Postdam

Universidad de Valladolid, *Institut de Mathématiques de Toulouse

2-class supervised classification

- **Objective** : forecast a label $Y \in \{0, 1\}$ using covariates $X \in \mathbb{R}^p$ using a classifier.
  A classifier is a function $g : \mathbb{R}^p \mapsto \{0, 1\}$ that predicts the label of an observation.
- Learning the classification rule from a learning sample.
  Observations : i.i.d copies $(Y_i, X_i) \in \{0, 1\} \times \mathbb{R}^p$ with $i = 1, \ldots, n$ of a random variable (Y, X) with distribution P.
- Defining the classification error : misclassification if $Y \neq g(X)$.

$$L(g) = 1_{\{Y \neq g(X)\}}$$

The error should be controled not only for learning sample but for all observations drawn with the same distribution.

$$R(g) = P((y, x) \in \{0, 1\} \times \mathbb{R}^p : y \neq g(x))$$

Best Classifier

$$f_\star = \arg\min_g P(Y \neq g(X))$$

is the best classifier (Bayes rule)

$$\eta(x) = P(Y = 1 | X = x)$$

$$f_\star(x) = 1_{\eta(x) \geq \frac{1}{2}}.$$

Not tractable : $\eta()$ is unknown since P is unknown.
Measure the difficulty of the problem

$$L^\star = L(f_\star).$$

Empirical Error vs Classification Error

$$R_n(g) := \frac{1}{n}\sum_{i=1}^{n} I_{(g(X_i)\neq Y_i)},$$

where $I_{(g(X)\neq Y)} = 1$ if $g(X) \neq Y$ and 0 otherwise.

- Select a class of classifier $\mathcal{F}$
- Optimize the classifier among the selected class

  $$\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} R_n(f).$$

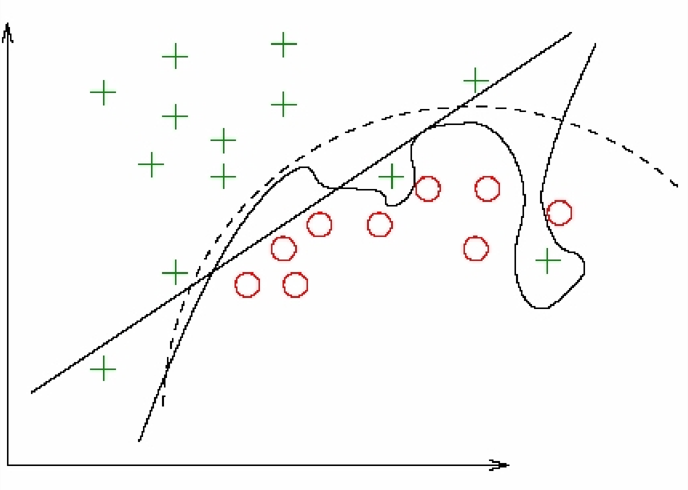
  Eventually control the complexity of the class to promote *sparsity*
- The classifier is evaluated not on the training set but for all similar observations

  $$L(\widehat{f}_n) = P(Y \neq \widehat{f}_n(X)).$$

Control the efficiency of the method with respect to optimal error

$$\mathcal{E}(\widehat{f}_n) = L(\widehat{f}_n) - L^\star.$$

Sometimes life is complicated



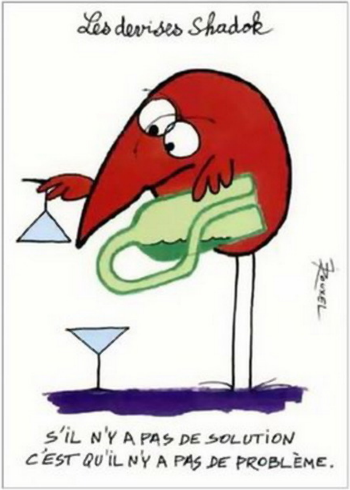
How to face difficulties ?

The error of a classifier depends on the classifier but on the distribution

$$P(Y \neq f(X)) = L(f, P).$$

- Trying to classify all points at all cost by boosting methods
  Constructing more complex classifiers or several classifiers and aggregating them ;
  Putting weights to the data which are badly classified and force the classifier to take them into account.
  A large amount of statistical literature ...
- .... or being (*pick you own adjective*) **optimistic**, ~~lazy~~, ~~data resilient~~, robust
  1. Accept to **say maybe** or **refuse to answer**
     Bartlett and Wegkamp (2014) (learning with reject option)
  2. Accept not to classify all points and remove some points ... of course not all : amounts to **change the distribution of the data**.

How to face difficulties ?



... the initial distribution of the data should not change too much

Removing data using trimming method

Definition 1

Given $\alpha \in (0,1)$, we define the set of α-trimmed versions of P by

$$\mathcal{R}_\alpha(P) := \left\{ Q \in \mathcal{P} :\ Q \ll P,\ \frac{dQ}{dP} \leq \frac{1}{1-\alpha}\ P - a.s. \right\}.$$

$Q \in \mathcal{R}_\alpha(P)$ can be seen as a close modification of a distribution P obtained by removing a certain quantity of data.
Given $\alpha \in (0,1)$, we define the trimmed classification error of a rule as the infimum of the α-trimmed probabilities of misclassifying future observations

$$R_\alpha(g) := \inf_{Q \in \mathcal{R}_\alpha(P)} Q(g(x) \neq y).$$

Admissible trimmed probabilities $P \equiv (p_0, P_0, P_1)$.

Let $A \subset \{0,1\} \times \mathbb{R}^p$, we denote $A_i = \{x \in \mathbb{R}^p : (i,x) \in A\}$, for $i = 0,1$.
$A = (\{0\} \times A_0) \cup (\{1\} \times A_1)$ $A \subset \{0,1\} \times \mathbb{R}^p$ and every probability $P \in \{0,1\} \times \mathbb{R}^p$,

$$P(A) = p_0 P_0(A_0) + p_1 P_1(A_1), \tag{1}$$

where $p_0 = P(\{0\} \times \mathbb{R}^p)$, $p_1 = 1 - p_0$,
$P_0(A_0) = P(A|Y=0) = P(\{0\} \times A_0)/p_0$ and
$P_1(A_1) = P(A|Y=1) = P(\{1\} \times A_1)/p_1$. P_0 P_1 probabilities in $\mathbb{R}^p$.

Lemma 2

$Q \equiv (q_0, Q_0, Q_1)$ with $q_0 \in (0,1)$, then $Q \in \mathcal{R}_\alpha(P)$ if and only if

$$q_0 \leq \frac{p_0}{1-\alpha}, \quad 1 - q_0 \leq \frac{1-p_0}{1-\alpha},$$

$$Q_0 \in \mathcal{R}_{1-\frac{q_0}{p_0}(1-\alpha)}(P_0) \quad \textit{and} \quad Q_1 \in \mathcal{R}_{1-\frac{1-q_0}{1-p_0}(1-\alpha)}(P_1).$$

How to minimize $Q \mapsto Q(g(x) \neq y)$?

$$Q(g(x) \neq y) = \int \left(q_0 I_{(g(x)=1)} \frac{dQ_0}{d\mu} + (1-q_0) I_{(g(x)=0)} \frac{dQ_1}{d\mu} \right) d\mu.$$

Aim : concentrate the probability Q_0 in the set $(g(x) = 0)$.

But $Q_0 \leq \frac{p_0}{q_0(1-\alpha)} P_0$

1. $P_0(g(x) = 0) \geq \frac{q_0}{p_0}(1-\alpha)$: As $\frac{p_0}{q_0(1-\alpha)} P_0 \geq 1$ we can group all the probability Q_0 in the set $\{x \in \mathbb{R}^p / g(x) = 0\}$ and hence $Q_0(g(x) = 0) = 1$.
2. $P_0(g(x) = 0) < \frac{q_0}{p_0}(1-\alpha)$: Now we can not give to $Q_0(g(x) = 0)$ probability 1, hence $Q_0(g(x) = 0) = \frac{P_0(g(x)=0)}{\frac{q_0}{p_0}(1-\alpha)}$.

Lemma 3

$$R_\alpha(g) = \min_{1-\frac{1-p_0}{1-\alpha} \leq q_0 \leq \frac{p_0}{1-\alpha}} \left[\left(q_0 - \frac{p_0}{1-\alpha} P_0(g(x) = 0) \right)_+ + \left(1 - q_0 - \frac{1-p_0}{1-\alpha} P_1(g(x) = 1)\right)_+\right.$$

Getting rid of all problems , if little problems ...

For fixed g, trimming reduces the classification error

Theorem 4

Given a trimming level $\alpha \in (0,1)$ *and a classification rule* g*,*

$$R_\alpha(g) = \frac{1}{1-\alpha}\left(R(g)-\alpha\right)_+ . \tag{2}$$

Recall $L^\star(Q) = \inf_g R_\alpha(g)$ achieved for Bayes classifier g_B^α

$$L_\alpha(P) := \inf_{Q \in \mathcal{R}_\alpha(P)} L^\star(Q) = \min_g R_\alpha(g) = R_\alpha(g_B^\alpha).$$

The following proposition compares these two errors.

$$\mathrm{Err}_\alpha(P) = \frac{(R(g_B)-\alpha)_+}{1-\alpha} = \frac{(L^\star-\alpha)_+}{1-\alpha}.$$

Empirical trimmed classification error

$$R(\mathcal{F}) := \min_{f \in \mathcal{F}} R(f) = R(f^\star).$$

In the same way we denote the trimmed error of the class $\mathcal{F}$ as $R_\alpha(\mathcal{F})$. Hence

$$R_\alpha(\mathcal{F}) := \min_{f \in \mathcal{F}} R_\alpha(f) = \min_{f \in \mathcal{F}} \frac{(R(f) - \alpha)_+}{1 - \alpha}.$$

Using empirical distribution :

$$R_{n,\alpha}(g) := \inf_{Q \in \mathcal{R}_\alpha(P_n)} Q(g(X) \neq Y)$$

$$R_{n,\alpha}(g) := \min_{w \in W} \sum_{j=1}^{n} w_i I_{(g(x_j) \neq y_j)} \tag{3}$$

with

$$W = \{w = (w_1, \ldots, w_n) /\ 0 \leq w_i \leq \frac{1}{n(1-\alpha)};\ i = 1, \ldots, n \wedge \sum_{i=1}^{n} w_i = 1\}.$$

Empirical Trimmed Classification Error

Empirical trimmed distribution reweighs the initial empirical distribution.

Controling the bias of empirical distribution :

Theorem 5

$$R_{n,\alpha}(g) = \frac{1}{1-\alpha}(R_n(g) - \alpha)_+.$$

$$0 \leq E(R_{n,\alpha}(g)) - R_\alpha(g) \leq \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)}.$$

A nice trick

Now let Y be a random variable such that $Y =^d X$, Y and X are independent, this implies $E(Y) = E_X(Y)$, using Jensen's inequality (for $(.)_+$) and conditional mean's properties we get

$$\begin{aligned} E((X - E(X))_+) &= E((X - E(Y))_+) = E((X - E_X(Y))_+) = E((E_X(X - Y \\ &\leq E(E_X((X - Y)_+)) = E((X - Y)_+). \end{aligned}$$

Now we are using that $X - Y$ is a symmetric variable, that it also is a centered variable,

$$\begin{aligned} E((X - Y)_+) &= \frac{1}{2}E(X - Y) \leq \frac{1}{2}(Var(X - Y))^{1/2} = \frac{1}{2}(Var(X) + Var(Y))^{1 \\ &= \frac{1}{2}(2Var(X))^{1/2}. \end{aligned}$$

How to select α ?

Trimmed models enable to decrease the classification error : **not sensitive to outliers or misclassified data**

Robust classification.

- **Selecting the amount of data to be removed**
- corresponds to the **optimal trimming level**.

Aim : Build a data driven rule to select $\hat{\alpha}$ which achieves balance between **minimization** of the classification risk without **removing a too large** quantity of information about the initial distribution.

For a fixed classifier : oracle inequality

Let $\xi_1 = (Y_1, X_1), \ldots, \xi_n = (Y_n, X_n)$ be n i.i.d with distribution P in $\{0,1\} \times \mathbb{R}^p$. Let g be a given classifier and $\alpha_{max} \in (0,1)$.

Theorem 6

Consider the penalization function

$$pen(\alpha) = \frac{1}{(1-\alpha)}\sqrt{\frac{\ln(n)}{2n}}$$

$$\hat{\alpha} = \arg\min_{\alpha \in [0,\alpha_{max}]} R_{n,\alpha}(g) + pen(\alpha),$$

then the following bound holds,

$$E(R_{\hat{\alpha}}(g)) \leq \inf_{\alpha \in [0,\alpha_{max}]} \left(R_\alpha(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{n}(1-\alpha)} \right)$$

$$+ \frac{1}{(1-\alpha_{max})}\sqrt{\frac{2\pi}{n}} + \frac{1}{n(1-\alpha_{max})^2}.$$

Proof

Typical proof of empirical risk minimization :

$$\hat{\alpha} = \arg \min_{\alpha \in [0, \alpha_{max}]} R_{n,\alpha}(g) + pen(\alpha)$$

$$R_{n,\hat{\alpha}}(g) + pen(\hat{\alpha}) \leq R_{n,\alpha}(g) + pen(\alpha).$$

$$R_{\hat{\alpha}}(g) \leq R_{\alpha}(g) + pen(\alpha) + (R_{n,\alpha}(g) - R_{\alpha}(g)) - pen(\hat{\alpha}) + (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)).$$

$$R_{n,\alpha}(g) - R_{\alpha}(g) = [R_{n,\alpha}(g) - E(R_{n,\alpha}(g))] + [E(R_{n,\alpha}(g)) - R_{\alpha}(g)],$$

Remains to control using a **concentration bound**

$$[R_{n,\alpha}(g) - E(R_{n,\alpha}(g))]$$

Proof

Mc Diarmid's inequality $R_{n,\alpha}(g) = F(\xi_1, \ldots, \xi_n)$ where $\xi_i = (Y_i, X_i)$. As

$$|F(\xi_1, \ldots, \xi_i, \ldots, \xi_n) - F(\xi_1, \ldots, \xi_i', \ldots, \xi_n)| \leq \frac{1}{n(1-\alpha)},$$

we can apply the inequality and hence

$$P(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \geq t) \leq e^{-2t^2 n(1-\alpha)^2}.$$

Given $z > 0$ take $t = \sqrt{\frac{z}{2n(1-\alpha)^2}}$, we get

$$P\left(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \geq \sqrt{\frac{z}{2n(1-\alpha)^2}}\right) \leq e^{-z}.$$

Proof

except in a set of probability not greater than e^{-z},

$$\begin{aligned} R_{\hat{\alpha}}(g) \leq R_{\alpha}(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} \\ + \sqrt{\frac{z}{2n(1-\alpha)^2}} - pen(\hat{\alpha}) \\ + (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)). \end{aligned}$$

$$R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) \leq \sup_{\alpha \in A}(R_{\alpha}(g) - R_{n,\alpha}(g))$$

Need for uniformity in Mc Diarmid achieved since α is in a compact set.

Extension to your favorite choice of classifiers

Let $\{\mathcal{G}_m\}_{m\in\mathbb{N}} \subset \mathcal{F}$ be a family of classes of classifiers with Vapnik-Chervonenkis dimension $V_{\mathcal{G}_m} < \infty$ for all $m \in \mathbb{N}$. Let $\alpha_{max} \in (0,1)$ and let Σ be a non-negative constant. Consider $\{x_m\}_{m\in\mathbb{N}}$ a family of non-negative weights such that

$$\sum_{m\in\mathbb{N}} e^{-x_m} \leq \Sigma < \infty.$$

Consider the penalization function

$$pen(\alpha, \mathcal{G}_m) = \sqrt{\frac{\ln(n) + x_m}{2n(1-\alpha)^2}} + \frac{1}{(1-\alpha)}\sqrt{\frac{V_{\mathcal{G}_m}\ln(n+1) + \ln(2)}{n}}$$

Define

$$(\hat{\alpha}, \hat{m}) = \arg \min_{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}} R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m),$$

Main Result

Theorem 7

$$E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \leq \min_{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} \right) + \frac{1+\Sigma}{2(1-\alpha_{max})}\sqrt{\frac{\pi}{2n}} + \frac{1}{n(1-\alpha_{max})^2}.$$

Need to know VC dimension of the classifiers.

Example of linear classifiers

$\mathcal{G}_m$ = family of linear classifiers built only using $x^{(m)}$ first m components of $X_i \in \mathbb{R}^p$.
Set $m \subset \mathcal{M} = \{1, \ldots, p\}$.

$$\mathcal{G}_m = \left\{ g \in \mathcal{F} : g(x) = I_{[a^T x^{(m)} + b \geq 0]}; a \in \mathbb{R}^m; b \in \mathbb{R} \right\}.$$

$$V_{\mathcal{G}_m} = m + 1.$$

We can choose $x_m = \ln(p)$ for all $m \in \mathcal{M}$ and $\Sigma = 1$.

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) &\leq \min_{(\alpha,m) \in [0,\alpha_{max}] \times \mathcal{M}} \left(R_\alpha(\mathcal{G}_m) + \sqrt{\frac{\ln(np)}{2n(1-\alpha)^2}} \right. \\ &+ \left. \frac{1}{(1-\alpha)} \sqrt{\frac{(m+1)\ln(n+1) + \ln(2)}{n}} + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} \right) \\ &+ \frac{1}{(1-\alpha_{max})} \sqrt{\frac{\pi}{2n}} + \frac{1}{n(1-\alpha_{max})^2}. \end{aligned}$$

good as long as $\ln(p)$ is smaller than n.

Comments : do not always trust the data

Removing data is not lazy point of view because sometimes the data are too numerous with many outliers (especially true in medicine)
The famous V **Veracity** in Big Data.

Convexity is easier for feasible minimization

Change the loss for a convex loss function

Hinge loss : $Y \in \{-1, 1\}$ $\gamma(x) = (1-x)_+$

$$L(Y, f(X)) = \gamma(Yf(X))$$

$$\begin{aligned} R(g) &= p_{-1} \int_0^{+\infty} P_{-1}(\{x \in \mathbb{R}^p : \gamma(-g(x)) \geq t\})dt \\ &+ p_1 \int_0^{+\infty} P_1(\{x \in \mathbb{R}^p : \gamma(-g(x)) \geq t\})dt. \end{aligned}$$

Theorem 8

Let $\alpha \in [0, 1)$,

$$R_\alpha(g) = \int_0^{\infty} \frac{(P(\{(y, x) : \gamma(yg(x)) > t\}) - \alpha)_+}{1-\alpha} dt. \quad (4)$$

As expected ...

Let $\xi_i = (Y_i, X_i)$ with $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^p$. Let $\{\mathcal{G}_m\}_{m\in\mathbb{N}}$ such that $V_{\mathcal{G}_m} < \infty$ for all $m \in \mathbb{N}$ and $|g(X_i)| \leq K$ with $K < +\infty$. $\alpha_{max} \in (0, 1)$

$$\sum_{m\in\mathbb{N}} e^{-x_m} \leq \Sigma < \infty.$$

$$pen(\alpha, \mathcal{G}_m) = \sqrt{\frac{2K^2(\ln(n) + x_m)}{n(1-\alpha)^2}} + \frac{2(1+K)}{1-\alpha}\sqrt{\frac{4V_{\mathcal{G}_m}\ln(n+1) + \ln(2)^3}{n\ln(2)^2}}$$

$$(\hat{\alpha}, \hat{m}) = \arg\min_{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}} R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m), \;\; g' := \arg\min_{g\in\mathcal{G}_m} R_\alpha(g),$$

Theorem 9

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \;\; & \leq \;\; \min_{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{1+K}{n(1-\alpha)} \right) \\ & + \;\; \frac{K(1+\Sigma)}{(1-\alpha_{max})}\sqrt{\frac{\pi}{2n}} + \frac{1+K}{n(1-\alpha_{max})}. \end{aligned}$$

In practice

Linear Regression : $g(X_i) = X_i\beta$

$$(\hat{\alpha}, \hat{m}) = \arg \min_{(\alpha,m)\in[0,\alpha_{max}]\times\{1,\dots,p\}} R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m)$$

$$= \arg \min_{(\alpha,m)\in[0,\alpha_{max}]\times\{1,\dots,p\}} \left[\min_W \sum_{i=1}^{n} w_i(1 - Y_i g(X_i))_+ + \sqrt{\frac{2K^2(\ln(n) + \ln(p))}{n(1-\alpha)^2}} \right.$$

$$\left. + \frac{2(1+K)}{1-\alpha}\sqrt{\frac{4V_{\mathcal{G}_m}\ln(n+1) + \ln(2)^3}{n\ln(2)^2}} \right]$$

where $W = \left\{ (w_1, \dots, w_n) : 0 \leq w_i \leq \frac{1}{n(1-\alpha)}; \sum w_i = 1 \right\}$
Iterative Algorithm to minimize

$$\min_W \sum_{i=1}^{n} w_i(1 - Y_i g(X_i))_+$$

In practice

Selecting $|H| = h = n - n\alpha$ points that provide the best residuals

$$\hat{\beta}_k = \arg\min_{\beta \in \mathbb{R}^m} Q(H_k, \beta) = \arg\min_{\beta \in \mathbb{R}^m} \sum_{i \in H_k} (1 - Y_i X_i \beta)_+ \quad (5)$$

with residuals

$$r_{ki} = (1 - Y_i X_i \hat{\beta}_k)_+.$$

Algorithm C-step and minimization with gradient descent.
Fit the regression on H_k then select the h smaller residuals among all observations then update H_{k+1}.
Application for medical data : remove fuzzy classified observations in large cohorts.

Extension : controlling the deviation using Wasserstein distance

$$R(Q) := Q(\{(y,x) : g(x) \neq y\}) + L\mathcal{W}_2^2(P_X, Q_X) \tag{6}$$

$$\begin{aligned} R_n(Q) = \min_{\pi_{i,j}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{ij} c_{ij} \\ \text{s.t} \quad & \sum_{j=1}^{n} \pi_{ij} \leq \frac{1}{n(1-\alpha)},\ i = 1, \ldots, n \\ & \sum_{i=1}^{n} \pi_{ij} = \frac{1}{n},\ j = 1, \ldots, n, \end{aligned}$$

with $c_{ij} = \ell(g; y_i, x_{ij}) + L \parallel X_i - X_j \parallel^2$ with $\ell(g; y_i, x_{ij})$ a positive loss function.

Need to study the statistical properties (open collaboration)

Extensions 2

Using regression

$$R_\alpha(g) = \inf_{Q \in \mathcal{R}_\alpha(P)} E_Q(\| Y - g(X) \|^2).$$

Let $F(t) = P(\{(y, x) : \| y - g(x) \|^2 \leq t\})$ and $\lambda = F^{-1}(1 - \alpha)$, so

$$R_\alpha(g) = \frac{1}{1 - \alpha}\left[R(g) - \lambda\alpha - \int_\lambda^\infty (1 - F(t))dt\right].$$

Possible to do the same things to get robust methods using Lasso penalty