Regression

Carlos C. Rodríguez
http://omega.albany.edu:8008/

November 1, 2004

Classical Regression

Recall the classical regression problem. Given observed data $(x_1, y_1), \ldots, (x_n, y_n)$ iid as a vector $(X, Y) \in R^{d+1}$ we want to estimate $f^*(x) = E(Y|X = x)$ i.e., the regression function of Y on X. When the vector (X, Y) is multivariate gaussian the regression function is $f^*(x) = \alpha + L(x)$ with $L(x)$ linear, and the Ordinary (yak!) Least Squares (OLS) estimator coincides with the MLE (Maximum Likelihood Estimator). Very often the distribution of (X, Y) is not explicitly known beyond the n observations and the available prior information about the meaning of these data. A typical assumption is to think of the y_j as the result of sampling the regression function at $f^*(x_j)$ with gaussian measurement error. The model is that conditionally on $x_1, \ldots, x_n$ the values of $y_1, \ldots, y_n$ are independent with Y_j depending on X_j only and $Y_j|X_j = x_j$ being $N(f^*(x_j), \sigma^2)$. Thus, for $j = 1, 2, \ldots, n$

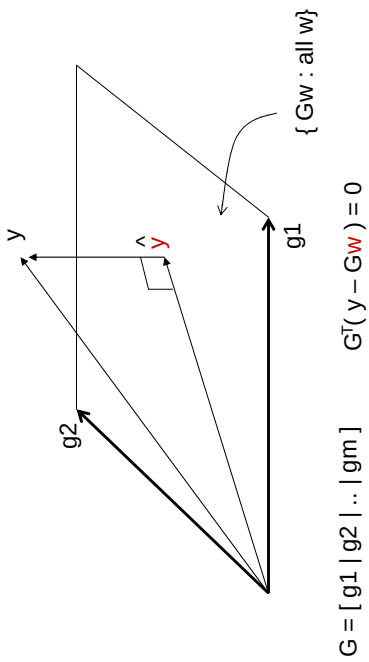
$$y_j = f^*(x_j) + \epsilon_j$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$. In this way the distribution of (X, Y) is modeled semiparametrically with (f, σ) where $f \in \mathcal{F}$ is a function in some space of functions $\mathcal{F}$ and $\sigma > 0$ is a positive scalar parameter. When $\mathcal{F}$ is taken as the m dimensional space $\mathcal{F}_m$ generated by functions, $g_1(x), \ldots, g_m(x)$ estimation of the regression function reduces to the linear optimization problem,

$$\hat{f} = \arg \min_{f \in \mathcal{F}_m} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

The solution is then given as the orthogonal (euclidean) projection of the observed vector $y^T = (y_1, y_2, \ldots, y_n)$ onto the space generated by the columns of the matrix $G \in R^{m \times n}$ with entries $g_{ij} = g_i(x_j)$. In fact, the above optimization problem can be written as,

$$||y - \hat{y}||^2 = \min_{w \in R^m} ||y - Gw||^2$$



As shown by the picture, the rejection vector (i.e. y minus its projection) must be orthogonal to the linear space generated by the columns of G, in particular (and equivalently) to each of these columns, obtaining the standard set of normal equations,

$$0 = G^T(y - \hat{y}) = G^T(y - G\hat{w})$$

with solution,

$$\hat{w} = (G^T G)^{-1} G^T y$$

The more general case of Weighted Least Squares (WLS) corresponding to the innerproduct $< x, z >= x^T A z$ generated by a symmetric positive definite matrix A, is just,

$$\hat{w} = (G^T A G)^{-1} G^T A y.$$

The matrix A encodes a covariance structure for the measurement errors, $\epsilon_1, \ldots, \epsilon_n$.

Over fitting and Kernel Regression

How should $\mathcal{F}$ be chosen?. On the one hand, we would like $\mathcal{F}$ to be big so not to constrain the form of the true regression function too much. On the other hand, big $\mathcal{F}$'s make the task of searching for the best $f \in \mathcal{F}$ more difficult and more importantly without a constraint on the explanatory capacity of $\mathcal{F}$ the solution will show no power of generalization. A big enough $\mathcal{F}$ will always have at least one member f, able to fit all the observations perfectly, without error, but this f provides no assurance that $f(x)$ is not as bad as it can possibly be for any point x not in the training set. To be able to assure that the size of the mistake on future data will not exceed a given value with high probability, (i.e. to have PAC bounds) we must constrain the capacity of $\mathcal{F}$ somehow. Over the years, statisticians and numerical analysts have invented all kinds of ad-hoc devices for achieving this goal. These are known as regularization methods. They boil down to adding a penalty term to the OLS empirical term, often of the form $\Omega(||f||)$ where Ω is an increasing function and $\mathcal{F}$ is assumed to be a space with a norm. The problem to be solved becomes,

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (y_i - f(x_i))^2 \text{ subject to } ||f|| \leq r_n$$

where the sequence of radiuses $r_n \to \infty$ as $n \to \infty$, but not too quickly, (at a given rate that depends of $\mathcal{F}$) so that some form of asymptotic stochastic convergence of the solution f_n towards the projection of the true regression function f^* onto $\mathcal{F}$ is achieved.

Kernel Regression

Reproducing Kernel Hilbert Spaces (RKHS) provide convenient choices for $\mathcal{F}$.

Theorem: Let K be a Mercer kernel and let $\mathcal{H}_K$ be the associated RKHS. If

$$C((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n)))$$

is any cost function that depends on $f \in \mathcal{H}_K$ only through the values of $f(x_j)$ at the observed x_j, then the minimizer of

$$U(f) = C\left((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))\right) + \Omega(||f||)$$

where Ω is an increasing function, is always achieved at a point $f_n \in \mathcal{H}_K$ of the form,

$$f_n(x) = \sum_{j=1}^{n} w_j K(x_j, x).$$

Thus, when $\mathcal{F} = \mathcal{H}_K$, a big fat infinite dimensional space, the regularized empirical cost $U = C + \Omega$ is minimized by solving a classic regression problem with $\mathcal{F}_n = \text{spann}\{K(x_1, \cdot), \ldots, K(x_n, \cdot)\}$.

Proof: The proof is surprisingly simple. Every $f \in \mathcal{H}_K$ can be written as $f = g + h$ where $g \in \mathcal{F}_n$ and $h \in \mathcal{F}_n^{\perp}$. We show that,

$$U(f) = U(g + h) \geq U(g)$$

with equality if and only if $h = 0$. This follows easily from the reproducing property of the kernel spaces. For all $j \leq n$,

$$f(x_j) = < K(x_j, \cdot), g + h > = < K(x_j, \cdot), g > = g(x_j)$$

since $h \perp \mathcal{F}_n$ by hypothesis. Thus, $C(f) = C(g)$. On the other hand, since Ω is strictly increasing and $g \perp h$, by the pythagorean theorem we have,

$$\begin{aligned} \Omega(||f||) &= \Omega\left((||g||^2 + ||h||^2)^{1/2}\right) \\ &\geq \Omega(||g||) \end{aligned}$$

with equality if and only if $h = 0$. Hence, $U(f) = C(g) + \Omega(||g + h||) \geq C(g) + \Omega(||g||) = U(g)$.•

Support Vector Regression

For given values $\alpha > 0$, and $\epsilon > 0$ define the empirical cost function ,

$$C = \alpha \sum_{i=1}^{n} |y_i - f(x_i)|_\epsilon$$

where,

$$|z|_\epsilon = \max\{0, |z| - \epsilon\}$$

is known as the ϵ insensitive function, and take,

$$\Omega(||f||) = \frac{1}{2}||f||^2.$$

With these choices, kernel regression becomes support vector regression. The parameter ϵ controls the sparness of the solution. The smoothing parameter α

controls the relative importance of the empirical cost C relative to the complexity penalty Ω.

The derivation of the support vector regression problem follows closely the derivation of support vector machines for classification. We first setup a primal optimization problem for minimizing the above ϵ-insensitive regularized empirical cost over functions, $f(x) =< w, x > +b$ for the euclidean innerproduct. Then we consider the dual problem. This turns out to be a simple quadratic programming problem that depends on the observed data only through the values of $(< x_i, x_j >)$. Just as in the classification case, we can apply the kernel trick and rip the benefits of nonlinear kernel regression at the linear regression cost!

The Primal Problem for SV Regression

We seek the solution of,

$$\text{minimize}\alpha \sum_{i=1}^{n} |y_i - f(x_i)|_\epsilon + \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

over b, w when $f(x) =< w, x > +b$. This is equivalent to,

$$\text{minimize}\alpha \sum_{i=1}^{n} u_i + \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

over u_i, w_i, b subject to: $u_i \geq |y_i - f(x_i)|_\epsilon$ for $i \leq n$. Each of the last n inequalities corresponds to three inequalities,

$$u_i \geq 0,\ u_i \geq y_i - f(x_i) - \epsilon,\ u_i \geq f(x_i) - y_i - \epsilon$$

Applying the standard trick of adding non negative slack variables ξ_i and ξ_i^* we soften the inequalities and allow small violations. So we replace the above constrained optimization problem with,

$$\text{minimize}\alpha \sum_{i=1}^{n} u_i + \frac{\alpha}{2}\sum_{i=1}^{n}(\xi_i + \xi_i^*) + \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

subject to: for $i \leq n$,

$$\begin{aligned} y_i - f(x_i) - \epsilon &\leq u_i + \xi_i \\ f(x_i) - y_i - \epsilon &\leq u_i + \xi_i^* \\ u_i \geq 0,\ \xi_i \geq 0,\ \xi_i^* &\geq 0. \end{aligned}$$

The objective function was chosen so that we can factorize out $\alpha/2$ and write,

$$\frac{\alpha}{2}\sum_{i=1}^{n}(\{u_i + \xi_i\} + \{u_i + \xi_i^*\}) + \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

In this way we can get rid of the u_i by just replacing $u_i + \xi_i$ by ξ_i and $u_i + \xi_i^*$ by ξ_i^* every where. Also, replace $\alpha/2$ by a new α to obtain the problem:

$$(Primal) \quad \text{minimize } \alpha \sum_{i=1}^{n} (\xi_i + \xi_i^*) + \frac{1}{2} \sum_{i=1}^{n} w_i^2$$

over, w_i, b, ξ_i, ξ_i^* subject to: for $i \leq n$,

$$\begin{aligned} & y_i - f(x_i) - \epsilon \leq \xi_i \\ & f(x_i) - y_i - \epsilon \leq \xi_i^* \\ & \xi_i \geq 0, \ \xi_i^* \geq 0 \\ & \text{and} \quad f(x) =< w, x > + b. \end{aligned}$$

The Dual Problem for SV Regression

The Lagrangian in terms of non negative Lagrange multipliers is,

$$\begin{aligned} \mathcal{L} \quad = \quad & \alpha \sum_i (\xi_i + \xi_i^*) + \frac{1}{2} \sum_i w_i^2 \\ & + \sum_i \lambda_i \{ y_i - f(x_i) - \epsilon - \xi_i \} \\ & + \sum_i \lambda_i^* \{ f(x_i) - y_i - \epsilon - \xi_i^* \} \\ & - \sum_i \beta_i \xi_i - \sum_i \beta_i^* \xi_i^*. \end{aligned}$$

To compute the dual we need to find,

$$W(\lambda, \lambda^*, \beta, \beta^*) = \min_{w,b,\xi,\xi^*} \mathcal{L}$$

The values of w, b, ξ, ξ^* where the minimum is achieved must satisfy,

$$\begin{aligned} \nabla_w \mathcal{L} \quad & = 0 \quad \Longleftrightarrow w = \sum_i (\lambda_i - \lambda_i^*) x_i \\ \nabla_b \mathcal{L} \quad & = 0 \quad \Longleftrightarrow \sum_i \lambda_i = \sum_i \lambda_i^* \\ \nabla_\xi \mathcal{L} \quad & = 0 \quad \Longleftrightarrow \lambda_j + \beta_j = \alpha \text{ for } j \leq n. \\ \nabla_{\xi^*} \mathcal{L} \quad & = 0 \quad \Longleftrightarrow \lambda_j^* + \beta_j^* = \alpha \text{ for } j \leq n. \end{aligned}$$

Replacing these equations into $\mathcal{L}$ we obtain that all the terms involving ξ_i and ξ_i^* dissapear from $\mathcal{L}$ and with them, β and β^*. Therefore, W is only a function

of λ and λ^*. We get, replacing $K(x_i, x_j)$ for the innerproducts $< x_i, x_j >$ (the kernel trick!) that,

$$\begin{aligned} W(\lambda, \lambda^*) &= \frac{1}{2}\sum_{i,j}(\lambda_i - \lambda_i^*)(\lambda_j - \lambda_j^*)K(x_i, x_j) \\ &\quad + \sum_i (\lambda_i - \lambda_i^*)y_i - \epsilon\sum_i(\lambda_i + \lambda_i^*) \\ &\quad - \sum_{i,j}(\lambda_i - \lambda_i^*)(\lambda_j - \lambda_j^*)K(x_i, x_j) \end{aligned}$$

The first and last terms simplify to produce,

$$\begin{aligned} W(\lambda, \lambda^*) &= -\epsilon\sum_i(\lambda_i + \lambda_i^*) \\ &\quad -\frac{1}{2}\sum_{i,j}(\lambda_i - \lambda_i^*)(\lambda_j - \lambda_j^*)K(x_i, x_j) \\ &\quad + \sum_i (\lambda_i - \lambda_i^*)y_i. \end{aligned}$$

The dual problem becomes,

$$(Dual) \quad \max_{\lambda, \lambda^*} W(\lambda, \lambda^*)$$

subject to:

$$\begin{aligned} &\sum_j \lambda_j = \sum_j \lambda_j^* \\ &0 \le \lambda_j \le \alpha,\ 0 \le \lambda_j^* \le \alpha. \end{aligned}$$

where we have replaced the equalities $\lambda_j + \beta_j = \alpha$, involving $\lambda_j \ge 0$, $\beta_j \ge 0$ by the equivalent inequalities shown above, that do not involve the βs.

As it was the case for classification, the dual problem is a simple quadratic programming problem that can be solved with efficient algorithms that are publicly available.

The solution from the QP solver is then used to produce the estimate,

$$\hat{w} = \sum_i (\lambda_i - \lambda_i^*)x_i$$

The KKT complementarity conditions, for the slack variables are of the type $\xi_j(\lambda_j - \alpha) = 0$ so we can write the other complementarity conditions as follows,

$$\begin{aligned} \lambda_j\{y_j - \hat{f}(x_j) - \epsilon\} &= 0 \quad \text{provided } \lambda_j < \alpha \\ \lambda_j^*\{\hat{f}(x_j) - y_j - \epsilon\} &= 0 \quad \text{provided } \lambda_j^* < \alpha \end{aligned}$$

are valid (and non trivial) for all $j \in J_0$, and $j \in J_0^*$ (resp.), where

$$J_0 = \{j : j \leq n, \text{ and } 0 < \lambda_j < \alpha\}$$

with J_0^* defined analogously.

These, and the complementarity conditions associated to the inequalities $\lambda_j \leq \alpha$ and $\lambda_j^* \leq \alpha$ make many $\lambda_j = \lambda_j^*$ to be either 0 or α and producing a sparse solution. The value for b can be obtained from any of the above complementarity conditions, but a more accurate value is obtained by combining efforts. Replacing the estimated values of the regression function at the training points,

$$\hat{f}(x_j) = \sum_i (\lambda_i - \lambda_i^*) K(x_i, x_j) + b$$

into the complementarity conditions, solving for b, multiplying through by λ_j and λ_j^* and adding over $j \in J$ with $J = J_0 \cap J_0^*$ we get,

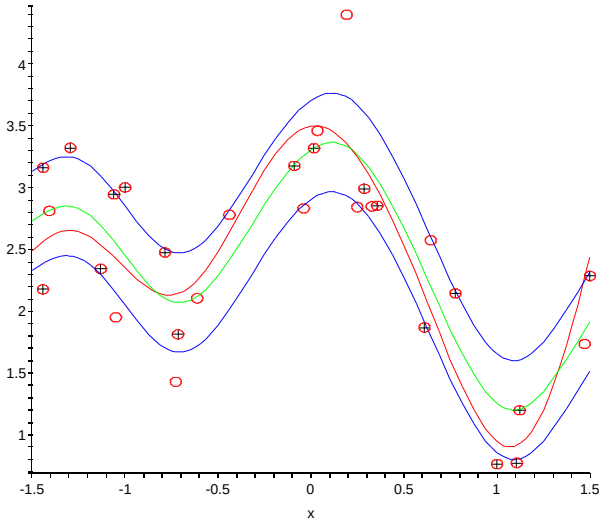
$$\begin{aligned} \sum_{j \in J} \lambda_j b &= \sum_{j \in J} \lambda_j \{y_j - \sum_i (\lambda_i - \lambda_i^*) K(x_i, x_j)\} \\ \sum_{j \in J} \lambda_j^* b &= \sum_{j \in J} \lambda_j^* \{y_j - \sum_i (\lambda_i - \lambda_i^*) K(x_i, x_j)\} \end{aligned}$$

adding the two equations, we finally obtain the estimate

$$b = \frac{\sum_{j \in J} (\lambda_j + \lambda_j^*) \{y_j - \sum_i (\lambda_i - \lambda_i^*) K(x_i, x_j)\}}{\sum_{j \in J} (\lambda_j + \lambda_j^*)}$$

Example: SV Regression in Action

The following picture shows $n = 30$ samples (the circles) from the true regression line (the red curve) with gaussian error and $\sigma = 0.5$. The green curve is the estimated regression line computed using a gaussian kernel. The blue curves show the $\epsilon = 0.4$ insensitive tube around the estimate. The support vectors are marked with plus signs and a value of $\alpha = 1.5$ was used. The Maple(9.5) code is available from this site and uses the QPsolve program in efficient matrix form from the new Maple optimization package.


4
3.5
3
2.5
2
1.5
1
-1.5
-1
-0.5
0
0.5
1
1.5
x