High Dimensional Inference and Regularization.

Eitan Greenshtein

November 17, 2006

Example

It is desired to train a machine to identify hand written digits for the purpose of recognizing hand written zip codes. The raw data, given to the machine, comes from $16 \times 16 = 256$ pixels. Denote the corresponding values $W_1, ..., W_{256}$.

In Vapnik (1998), a construction of a classifier is described, which is a function of all 'interactions' up to order 7 of $W_1, ..., W_{256}$.

This creates $p \approx 10^{16}$, explanatory variables, $X_1, ..., X_p$. They had $n = 7291$ examples (or data points).

They were searching for a good linear classifier, i.e., one that classifies to one of two candidate groups based on the sign of

$$\beta_0 + \sum \beta_j X_j,$$

for 'appropriate' $\beta_0, \beta_1, ..., \beta_p$.

How to select 'appropriate' $\beta_0, ..., \beta_p$??

Regularization

Consider the equation

$$Y = X\beta,$$

where $Y' = (Y_1, ..., Y_n)$ is an n dimensional vector, $X = (X_{ij})$ is a $n \times p$ matrix, and $\beta' = (\beta_1, ..., \beta_p)$.

We want to solve for β.

Suppose, X is non-singular.

Then

$$\beta = (X'X)^{-1}X'Y \ (= X^{-1}Y).$$

Note: the solution may be very 'unstable' if $(X'X)$ is close to singular.

Unstable means: small changes in Y_i could change dramatically the solution β.

Consider the integral equation

$$y(x) = \int_0^1 K(x,t)f(t)dt,$$

where we want to solve for $f(t)$.

When we consider only a grid of points, we get analogous of system of linear equations as discussed above.

Solving it numerically on a grid of points, could again produce very unstable solutions, i.e., a slight numerical inaccuracy in the values of the grid points would dramatically affect the solution when the number of grid points is large.

Given a grid of points, Tikhonov suggested, rather than finding the exact solution, to find a close solution which is "regularized". Regularized means satisfying some constraints, on its smoothness or its norm, etc.

In matrix/linear-equation form, an example of a regularization method would be,

$$argmin_\beta ||Y - X\beta||^2 + \lambda(\sum_j \beta_j^2), \;\; \lambda \geq 0.$$

The right choice of λ in relation to the accuracy and size of the grid, was studied by Tikhonov.

This particular method of regularization was also suggested in statistics and called Ridge-Regression.

A method that is extensively studied in Statistics last 10 years is:

$$argmin_\beta ||Y - X\beta||^2 + \lambda \sum_j |\beta_j|, \;\; \lambda \geq 0.$$

called LASSO.

In statistics we have a similar situation.

In classical regression, we want to find β which is the solution of

$$E(Y) = X\beta.$$

We have 'inaccuracy' since we observe Y rather than $E(Y)$.

We are also interested in a 'stable' solution.

Vapnik, suggested that some of the ideas of Tikhonov should be relevant for statistics, and we should also look for the right regularization methods. For example, in high dimensional regression, do not just find least squares, but impose regularization in order to get 'stable' solutions.

What is meant by a 'stable' solution, in a statistical context?

Suggestions and Heuristics

Consider a regression problem. We want to find a 'good' linear predictor for Y, based on $X_1, ..., X_p$.

'Good'- e.g., has a low value of $E(Y - \sum \beta_j X_j)^2$.

Given data, Y_i and corresponding $X_{i1}, ..., X_{ip}$, $i = 1, ..., n$, we write the corresponding vector and matrix by Y^D and X^D.

Let

$$\hat{\beta} = argmin_\beta ||Y^D - X^D\beta||^2 + regularization-penalty.$$

A solution/estimator $\hat{\beta}$ is stable if

$$E||Y - X\hat{\beta}||^2 \approx ||Y^D - X^D\hat{\beta}||^2.$$

Regression/Prediction with Random Explanatory Variables

We have an independent random sample,
$Z_i = (Y_i, X_{i1}, ..., X_{ip})'$, $Z_i \sim F$, $i = 1, .., n$.

We want to find $\beta' = (\beta_1, ..., \beta_p)$, so that $\sum \beta_j X_j$ is a good predictor for the corresponding Y.

A good predictor, in the sense (e.g.)

$$E_F(Y - \sum_{j=1}^{P} \beta_j X_j)^2$$

is small.

Denote $\alpha' = (-1, \beta_1, ..., \beta_p)$ and $\beta' = (\beta_1, ..., \beta_p)$.

Note:

$$E_F(Y - \sum_{j=1}^{P} \beta_j X_j)^2 = \alpha' \Sigma \alpha,$$

where $\Sigma = E_F Z Z'$.

Given $(X_1, ..., X_p)$, let

$$U = E(X_1, ..., X_p)'(X_1, ..., X_p).$$

Hence, $U_{kl} = EX_kX_l$.

Let $V' = (V_1, ..., V_n)$, where $V_j = EX_jY$.

If U is non-singular, then the vector β^* which minimizes

$$E_F(Y - \sum_{j=1}^{P} \beta_j X_j)^2 = \alpha'\Sigma\alpha,$$

is:

$$\beta^* = U^{-1}V.$$

Note: neither U nor V are known.

A naive practice is to replace U and V by their empirical version denoted $\hat{U}$ and $\hat{V}$.

Here, $\hat{U}_{kl} = \frac{\sum_i X_{ik} X_{il}}{n}$, $\hat{V}_j = \frac{\sum_i Y_i X_{ij}}{n}$.

($\hat{U}^{-1}\hat{V} = (X'X)^{-1}X'Y$ is the familiar least squares estimator).

Note!, when n is large, by the law of large numbers, for each k, l, $\hat{U}_{kl}$, converges to U_{kl}. Yet, when p is comparable or larger than n, U and $\hat{U}$ are not close in a matrix norm, i.e., as operators. Similar considerations apply for the vectors $\hat{V}$ and V.

Thus, special care is required in estimation of such high dimensional matrices and vectors.

We should apply appropriate regularization methods, in order to obtain 'stable solutions' to the problem of interest.

A special interest and effort is in 'sparse' situation, where most entries are (nearly) zero.

This is expected in data mining applications.

Example of Estimation of a High Dimensional Vector of Means.

The sequence $\mu_1, \mu_2, ..., \mu_n$ of Binary signals is transmitted. It is corrupted and the receiver 'observe' $Y_1, Y_2, ...Y_n$, where $Y_i \sim N(\mu_i, 1)$ are independent.

The 'commonsense' method to recover the sequence is by:
$\hat{\mu}_i = 1$ if $\phi(Y_i - 1) > \phi(Y_i)$, $\hat{\mu}_i = 0$ otherwise.

This is also the m.l.e. when $(\mu_1, ..., \mu_n)$ is confined to be a binary vector.

Let p be the proportion of 1's in the sequence. Then $\frac{\sum Y_i}{n} = \hat{p} \approx p$.

Consider the recovery method:
$\tilde{\mu}_i = 1$ if $\hat{p}\phi(Y_i - 1) > (1 - \hat{p})\phi(Y_i)$, $\tilde{\mu}_i = 0$ otherwise.

Under 0-1 loss, as $n \to \infty$, for $p \neq 0.5$ the second method will dominate the first one.

Under a squared error loss the same reasoning suggests using the estimator:

$$\hat{\delta}_1(Y_i) = \frac{\hat{p}\phi(Y_i - 1)}{\hat{p}\phi(Y_i - 1) + (1 - \hat{p})\phi(Y_i)}.$$

Denote:

$$\delta_1(Y_i) = \frac{p\phi(Y_i - 1)}{p\phi(Y_i - 1) + (1 - p)\phi(Y_i)}.$$