Regularized Least Squares to Remove Reviewer Bias

by John Platt and Chris Burges

May 2012

Given a set of assigned papers, some reviewers may have a tendency to give high scores, and some low. This can throw off the overall quality of the reviewing. Here we present the model that C.B. and Léon Bottou, co-Program Chairs for Neural Information Processing Systems 2012, used to estimate the reviewer bias for that conference. This model was introduced much earlier, by J.P. for NIPS 2002, and it, or a variant of it, has been used at many subsequent NIPS conferences. The model is similar to a two-way ANOVA model.

These bias estimates are mostly intended to help Area Chairs (who, for NIPS, each typically handle reviews for 20 to 30 papers) identify overly positive or negative reviewers and to adjust their decisions accordingly. For example, if an Area Chair notices that all reviewers for a paper agree well except for reviewer X, then that AC can check X's bias score to see if this helps explain the difference. In NIPS 2012, the use of this data by the ACs was optional – it was just intended as an aid. C.B. decided to provide the description below, together with a computation showing how the model avoids overfitting, for those who were curious as to how these scores were computed, and in case organizers of other conferences might find this bias detection method useful.

A reviewer bias plus paper quality regularized model

This model estimates the per-reviewer biases and a per-paper quality score. It also uses a regularization term to handle reviewers with small numbers of assigned papers. One might worry that adding a parameter for each paper, as well as for each reviewer (1404 extra parameters, for NIPS 2012!) could lead to overfitting. We show in the section below that this is not the case – a similar model that models only the per-reviewer biases arrives at the same results for those biases.

We minimize the regularized least-squares:

$$L = \frac{1}{2}\sum_{ij} A_{ij}(r_{ij} - g_i - b_j)^2 + \frac{\lambda}{2}\sum_j b_j^2$$

where i indexes papers, j indexes reviewers, and A_{ij} is a sparse matrix of 0's and 1's indicating that reviewer j scored paper i. Here r_{ij} is the score reviewer j gave to paper i, g_i is the corrected score for paper i (call it the 'goodness' of paper i), b_j is the bias of reviewer j, and λ is the regularization parameter. Note that the regularization term pulls the biases towards zero.

Setting the gradients to zero gives:

$$\frac{\partial L}{\partial g_i} = -\sum_j A_{ij}(r_{ij} - g_i - b_j) = 0 \qquad (1)$$

$$\frac{\partial L}{\partial b_j} = -\sum_i A_{ij}(r_{ij} - g_i - b_j) + \lambda b_j = 0 \qquad (2)$$

which gives

$$g_i = \frac{\sum_j A_{ij}(r_{ij} - b_j)}{\sum_j A_{ij}}$$
$$b_j = \frac{\sum_i A_{ij}(r_{ij} - g_i)}{\sum_i A_{ij} + \lambda} .$$

Note that the goodness g_i is simply the average of the reviewer scores, with each score corrected with the bias of the reviewer, namely $r_{ij} - b_j$. Thus for $\lambda = 0$, Eq. (2) can be written

$$\sum_i A_{ij}(r_{ij} - b_j) = \sum_i A_{ij} g_i \qquad (3)$$

which is simply saying that the sum of the jth reviewer's scores for their papers must equal the sum of the average (corrected) scores for those papers. We will find below that the simpler reviewer-bias-only model gives exactly the same result.

Note also that the regularization term behaves as if there are λ extra reviews for each reviewer, for each of which the reviewer's score is equal to the paper's goodness. This means that if a reviewer provides fewer than λ reviews, then their bias estimate is strongly pulled towards zero. This interpretation provides an intuitive way to set λ.

Solving for g_i and b_j therefore amounts to just solving a linear system. Let

$$x_i = \sum_j A_{ij}, \quad y_j = \sum_j A_{ij} + \lambda, \quad s_i = \sum_j A_{ij} r_{ij} = \sum_j r_{ij}, \quad t_j = \sum_i A_{ij} r_{ij} = \sum_i r_{ij}$$

Then, the linear system (which has dimension number of reviewers plus number of papers) is:

$$\begin{bmatrix} \boldsymbol{X_i} & \boldsymbol{A} \\ \boldsymbol{A^T} & \boldsymbol{Y_j} \end{bmatrix} \begin{pmatrix} g_i \\ b_j \end{pmatrix} = \begin{pmatrix} s_i \\ t_j \end{pmatrix}$$

where $\boldsymbol{X_i}$ ($\boldsymbol{Y_j}$) is the matrix with the x_i (y_j) along the diagonal, zeros elsewhere. The problem is well-conditioned, due the regularization.

Note that this formulation also permits the use of the reviewer confidences: A_{ij} takes the form of a precision of a Gaussian. Instead of 1 for all papers, we can set it to be the confidence that reviewer j has in his review of paper i. This makes the innate goodness the weighted average of the corrected reviews. (However we did not used this feature in the numbers we provided for NIPS 2012.)

The bias-only model

Here we show that including the per-paper 'goodness' fits does not lead to overfitting, in the sense that solving the problem directly for reviewer biases alone gives the same result for the reviewer biases (with $\lambda = 0$).

Again let r_{ij} be the score that reviewer j assigns to paper i and let the bias for reviewer j be b_j so that the adjusted score is $r_{ij} - b_j$. The mean adjusted score for paper i is

$$\mu_i := \frac{1}{|R_i|} \sum_{j \in R_i} (r_{ij} - b_j)$$

where R_i is the set of reviewers for paper i. The un-normalized variance is

$$\sigma_i^2 = \sum_{j \in R_i} (r_{ij} - b_j - \mu_i)^2 = \sum_{j \in R_i} (r_{ij} - b_j)^2 - |R_i| \mu_i^2$$

and the total variance we wish to minimize is $\sum_i \sigma_i^2$.

Let T_j be the set of papers assigned to reviewer j. Then for an arbitrary reviewer k,

$$\begin{aligned}
\frac{\partial \sum_i \sigma_i^2}{\partial b_k} &= \sum_{i \in T_k} \frac{\partial}{\partial b_k} \left(\sum_{j \in R_i} (r_{ij} - b_j)^2 - |R_i| \mu_i^2 \right) \\
&= \sum_{i \in T_k} \sum_{j \in R_i} -2(r_{ij} - b_j)\delta_{jk} - \sum_{i \in T_k} \frac{\partial}{\partial b_k} \frac{1}{|R_i|} \sum_{m \in R_i} \sum_{n \in R_i} (r_{im} - b_m)(r_{in} - b_n) \\
&= \sum_{i \in T_k} -2(r_{ik} - b_k) + \sum_{i \in T_k} \frac{2}{|R_i|} \sum_{j \in R_i} (r_{ij} - b_j) = 0 \ .
\end{aligned}$$

This is simply saying that for the kth reviewer, the sum of her scores for her papers must equal the sum of mean scores for those papers. This is the same condition found in the previous analysis. To write this in matrix form, let's use the same indicator function as before: let $A_{ak} = 1$ if reviewer k has paper a, 0 otherwise. Then we can write the condition for optimality as

$$\sum_a (r_{ak} - b_k) A_{ak} = \sum_{a,i} \frac{1}{|R_a|} (r_{ai} - b_i) A_{ai} A_{ak}$$

Define the scores vector

$$S_k := -\sum_a r_{ak} A_{ak} + \sum_{a,i} \frac{1}{|R_a|} r_{ai} A_{ai} A_{ak} \ .$$

Also note that $\sum_a A_{ak} = |T_k|$. Let $\Pi_{ai} := \frac{A_{ai}}{|R_a|}$. Thus for each k, we must have

$$-|T_k| b_k + \left(\mathrm{A}^{\mathrm{T}} \Pi b\right)_k = S_k \ .$$

Hence if T is the diagonal matrix whose kth diagonal entry is $|T_k|$, we just have to solve the linear system

$$\left(\mathrm{A}^{\mathrm{T}} \Pi - T\right) b = S_k \, .$$

Note that this is a smaller problem than the previous one (number of reviewers × number of reviewers).