Reason for not shrinking the bias (intercept) term in regression

For linear model, $y = \beta_0 + x\beta + \varepsilon$, the shrinkage term is always $P(\beta)$.

What is the reason that we do not shrink the bias (intercept) term β_0? Should we shrink the bias term in the neural network models?

regression | neural-networks | ridge-regression | intercept | shrinkage

edited May 14 '16 at 23:00
amoeba
40.3k 10 138 203

asked Feb 18 '14 at 9:50
yliueagle
318 3 10

The liblinear library for logistic regression as used in scikit-learn penalises the bias term (I think this is an implementation artifact, bias is handled as extra input variable) – seanv507 Sep 28 '15 at 0:05

4 Answers

The Elements of Statistical Learning by Hastie et al. define ridge regression as follows (Section 3.4.1, equation 3.41):

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

i.e. explicitly exclude the intercept term β_0 from the ridge penalty.

Then they write:

> [...] notice that the intercept β_0 has been left out of the penalty term. Penalization of the intercept would make the procedure depend on the origin chosen for Y; that is, adding a constant c to each of the targets y_i would not simply result in a shift of the predictions by the same amount c.

Indeed, in the presence of the intercept term, adding c to all y_i will simply lead to β_0 increasing by c as well and correspondingly all predicted values $\hat{y}_i$ will also increase by c. This is not true if the intercept is penalized: β_0 will have to increase by less than c.

In fact, there are several nice and convenient properties of linear regression that depend on there being a proper (unpenalized) intercept term. E.g. the average value of y_i and the average value of $\hat{y}_i$ are equal, and (consequently) the squared multiple correlation coefficient R is equal to the coefficient of determination R^2:

$$(R)^2 = \cos^2(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\|\hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2} = R^2,$$

see e.g. this thread for an explanation: Geometric interpretation of multiple correlation coefficient R and coefficient of determination R^2.

Penalizing the intercept would lead to all of that not being true anymore.

edited Apr 13 at 12:44
Community ♦
1

answered Jul 15 '15 at 23:30
amoeba
40.3k 10 138 203

Recall the purpose of shrinkage or regularization. It is to prevent the learning algorithm to overfit the training data or equivalently - prevent from picking arbitrarily large parameter values. This is more likely for datasets with more than few training examples in the presence of noise (very interesting discussion about presence of noise and its impact is discussed in "Learning from Data" by Yaser Abu-Mustafa). A model learned on noisy data with no regularization will likely perform poorly on some unseen data points.

With this in mind, imagine you have 2D data points which you want to classify in two classes. Having all but the bias parameters fixed, varying the bias term will just move the boundary up or down. You can generalize this to a higher dimensional space.

The learning algorithm cannot put arbitrarily large values for the bias term since this will result in possibly gross loss value (the model will not fit the training data). In other words, given some training set, you (or a learning algorithm) cannot move the plane arbitrarily far away from the true one.

So, there is no reason to shrink the bias term, the learning algorithm will find the good one without a risk of overfitting.

A final note: I saw in some paper that when working in high-dimensional spaces for classification, there is no strict need to model the bias term. This might work for linearly separable data since with more dimensions added, there are more possibilities to separate the two classes.

edited Jul 15 '15 at 23:53

answered Jul 15 '15 at 23:32

xeon
1,107 6 12

Can you give references for some papers which says "when working in high-dimensional spaces for classification, there is no strict need to model the bias term"? – chandresh Feb 2 '16 at 12:09

The intercept term is absolutely not immune to shrinkage. The general "shrinkage" (i.e. regularization) formulation puts the regularization term in the loss function, e.g.:

$$RSS(\beta) = ||y_i - X_i\beta||^2$$

$$RegularizedLoss(\beta) = RSS(\beta) - \lambda f(\beta)$$

Where $f(\beta)$ is usually related to a lebesgue norm, and λ is a scalar that controls how much weight we put on the shrinkage term.

By putting the shrinkage term in the loss function like this, it has an effect on *all* the coefficients in the model. I suspect that your question arises from a confusion about notation in which the β (in $P(\beta)$) is a vector of all the coefficients, inclusive of β_0. Your linear model would probably be better written as $y = X\beta + \epsilon$ where X is the "design matrix," by which I mean it is your data with a column of $1's$ appended to the left hand side (to take the intercept).

Now, I can't speak to regularization for neural networks. It's possible that for neural networks you want to avoid shrinkage of the bias term or otherwise design the regularized loss function differently from the formulation I described above. I just don't know. But I strongly suspect that the weights and bias terms are regularized together.

edited Feb 18 '14 at 16:28

answered Feb 18 '14 at 13:47

David Marx
4,602 12 34

2 It depends on the convention, but e.g. The Elements of Statistical Learning by Hastie et al. define ridge regression such that intercept is not penalized (see my answer). I suspect this might be more standard than otherwise. – amoeba Jul 15 '15 at 23:32

I'm not sure the above answer by David Marx is quite right; according to Andrew Ng, by convention the bias/intercept coefficient is typically not regularized in a linear regression, and in any case whether it is regularized or not does not make a significant difference.

answered Jul 15 '15 at 23:11

xenocyon
111 3