Machine Learning Introduction

Regularization and bias variance trade-off

Blanca Cano Camarero

Universidad Autónoma de Madrid

11 November 2022

Last meet

- What is machine learning
- Problem abstraction
- Lineal models

This week

- Quantil regression
- Regularization
- Bias variance trade off

About ten hours of work.

The troubles of expectance

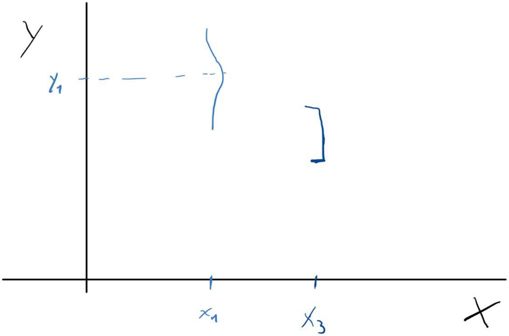


Figure 1: Troubles

Expectance approximation

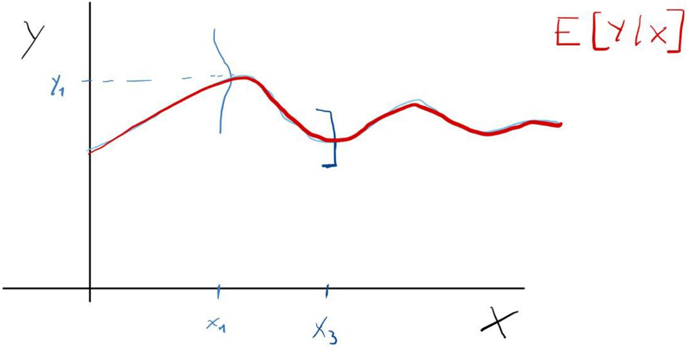


Figure 2: Expectance approach

Quantil approach

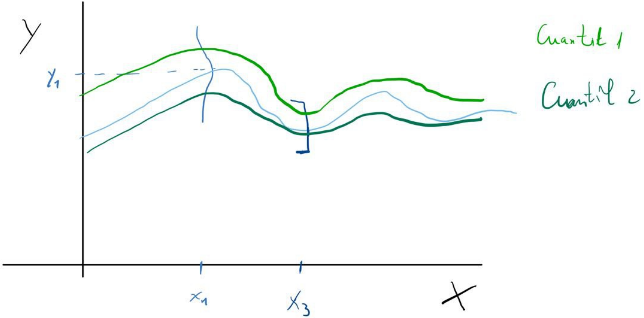
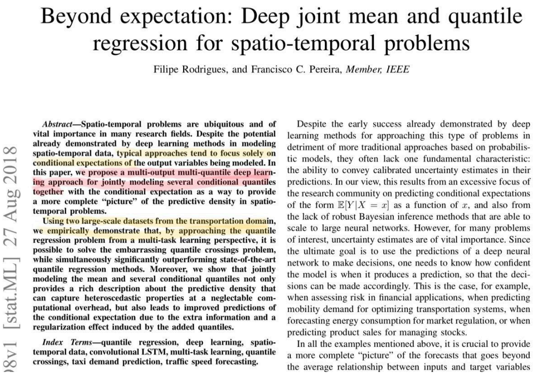


Figure 3: Quantil approach

Article



Beyond expectation: Deep joint mean and quantile regression for spatio-temporal problems

Filipe Rodrigues, and Francisco C. Pereira, *Member, IEEE*

98v1 [stat.ML] 27 Aug 2018

Abstract—Spatio-temporal problems are ubiquitous and of vital importance in many research fields. Despite the potential already demonstrated by deep learning methods in modeling spatio-temporal data, typical approaches tend to focus solely on conditional expectations of the output variables being modeled. In this paper, we propose a multi-output multi-quantile deep learning approach for jointly modeling several conditional quantiles together with the conditional expectation as a way to provide a more complete "picture" of the predictive density in spatio-temporal problems.

Using two large-scale datasets from the transportation domain, we empirically demonstrate that, by approaching the quantile regression problem from a multi-task learning perspective, it is possible to solve the embarrassing quantile crossings problem, while simultaneously significantly outperforming state-of-the-art quantile regression methods. Moreover, we show that jointly modeling the mean and several conditional quantiles not only provides a rich description about the predictive density that can capture heteroscedastic properties at a neglectable computational overhead, but also leads to improved predictions of the conditional expectation due to the extra information and a regularization effect induced by the added quantiles.

Index Terms—quantile regression, deep learning, spatio-temporal data, convolutional LSTM, multi-task learning, quantile crossings, taxi demand prediction, traffic speed forecasting.

Despite the early success already demonstrated by deep learning methods for approaching this type of problems in detriment of more traditional approaches based on probabilistic models, they often lack one fundamental characteristic: the ability to convey calibrated uncertainty estimates in their predictions. In our view, this results from an excessive focus of the research community on predicting conditional expectations of the form $\mathbb{E}[Y|X = x]$ as a function of x, and also from the lack of robust Bayesian inference methods that are able to scale to large neural networks. However, for many problems of interest, uncertainty estimates are of vital importance. Since the ultimate goal is to use the predictions of a deep neural network to make decisions, one needs to know how confident the model is when it produces a prediction, so that the decisions can be made accordingly. This is the case, for example, when assessing risk in financial applications, when predicting mobility demand for optimizing transportation systems, when forecasting energy consumption for market regulation, or when predicting product sales for managing stocks.

In all the examples mentioned above, it is crucial to provide a more complete "picture" of the forecasts that goes beyond the average relationship between inputs and target variables

See Rodrigues and Pereira (2018)

Quantile regression

Instead of using $E[Y|X]$ change it by quantiles See "Quantile Regression" (2022) and "Quantile Regression in Machine Learning" (2018)

Problems

- When there are many correlated variables in a linea regression model, their coefficient can became poorly determined and exhibit high variance.
- Wild large positive coefficient on one variable can be canceled y a similarly large positive coefficient on its correlated cousin.
- Model according to the size of the available training set to avoid over fitting.
- The number of parameters in not necessarily the most appropriate measure of model complexity.

Solution

- Retaining a subset of the predictions of discarding the rest exhibits high variance.
- By imposing a size constraint on the coefficient this problem is alleviated.
- We often would like to determine a smaller subset that exhibit the strongest effect.

Subset Selection

Find the best subset of size k that gives smallest residual sum of squares.

And efficient algorithm is *leap and bounds* (1975)

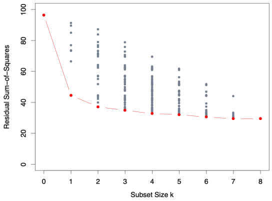


FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Figure 4: Best subset example from section 3.3.1 Hastie, Tibshirani, and Friedman (2001)

Shrinkage methods

- Shrink are more continuous and don't suffer as much from high variability.

Shrinkage methods: Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size.
- The ridge coefficients minimize a penalized residual sum of

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

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Ridge regression properties

- There is a one to one correspondence between the parameter λ and t.

An equivalent way to write the ridge problem is

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

Where $\lambda \geq 0$ is a complexity parameter t that controls the amount of shrinkage.

Matrix form

Denote by X the $X \times (p+1)$ matrix with each row an input vector (with a 1 in the first position)

$$RSS(\lambda) = (y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

The ridge regression solution are easily seen to be

$$\hat{\beta}^{\text{ridge}} = (X^TX + \lambda I)^{-1}X^TY. \tag{3}$$

Proof ridge

$$RSS(\lambda) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \tag{4}$$

$$\frac{\partial RSS}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta \tag{5}$$

$$\frac{\partial\partial RSS}{\partial\beta\partial\beta^T} = 2X^TX + 2\lambda \tag{6}$$

Proof

Assuming that X has full column rank, hence X^TX is positive definite, and $\lambda > 0$ we set the first derivative to zero

$$X^T(y - X\beta) = \lambda\beta \tag{7}$$

to obtain the unique solution

$$\hat{\beta} = (X^TX - I\lambda)^{-1}X^TY. \tag{8}$$

Which components are more affected by shrinkage ?

Singular value decomposition

$$X = UDV^T \tag{9}$$

Here U and V are $N \times p$ and $p \times p$ orthogonal matrices.

D is a diagonal matrix of singular values :

$$d_1 \geq d_2 \geq ... \geq d_p \geq 0$$

.

Why D is Positive semi-definite? (Defines an inner product, see “Matriz Definida Positiva” (2022) and “Producto Escalar” (2022))

Geometrical understanding

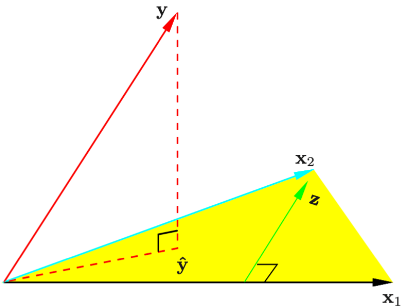


FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector $\mathbf{x}_2$ is regressed on the vector $\mathbf{x}_1$, leaving the residual vector $\mathbf{z}$. The regression of $\mathbf{y}$ on $\mathbf{z}$ gives the multiple regression coefficient of $\mathbf{x}_2$. Adding together the projections of $\mathbf{y}$ on each of $\mathbf{x}_1$ and $\mathbf{z}$ gives the least squares fit $\hat{\mathbf{y}}$.*

Figure 5: Geometrica understanding

Singular value decomposition

Now the ridge regression is

$$\begin{aligned} X\hat{\beta}^{\text{ridge}} &= X(X^TX - I\lambda)^{-1}X^TY && (10)\\ &= UD(D^2 + \lambda I)^{-1}DU^TY && (11)\\ &= \sum_{j=1}^{p} u_{*i}\frac{d_{jj}^2}{d_{jj}^2 + \lambda}u_{*j}^T y_j. && (12)\end{aligned}$$

Lasso

$$\hat{\beta}^{\text{lasso}} = \text{argmin}_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{13}$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Lasso in the equivalent *Lagragian form*

$$\hat{\beta}^{\text{lasso}} = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (14)$$

Also known as a **sparse model** since some coefficient converge to zero.

Computing the lasso solution is a quadratic programming (see "Quadratic Programming" (2022)) Least Angle Regression

Estimation picture for laso and ridge regression

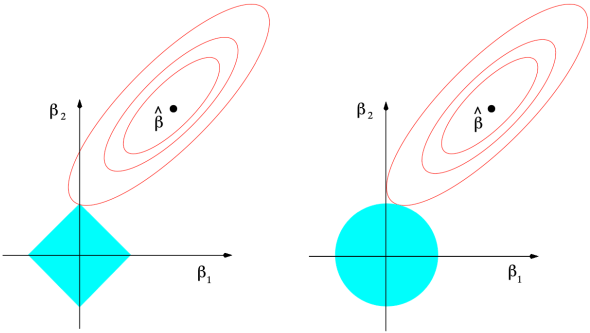


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Contours for given q

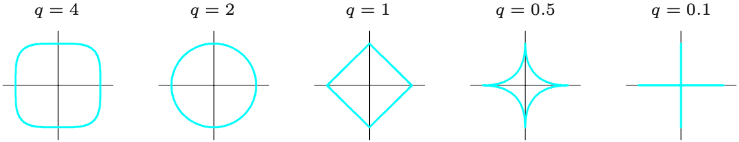


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Figure 7: Contours for given q

Elastic net penalty

$$\lambda \sum_j \alpha \beta_j^2 + (1-\alpha)|\beta_j|$$

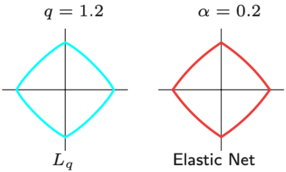


FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j(\alpha\beta_j^2+(1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

Other generalizations ?

Other generalizations ?

- Elastic net penalty generalization
- Introducing prior knowledge

More generalized

$$\hat{\beta}^{\text{general}} = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \alpha_j |\beta_j|^p \right\} \tag{15}$$

subject to $\alpha_j \geq 0$ and

$$\sum_{j=1}^{p} \alpha_j = 1.$$

Cons: It is worthy since this method is an heuristic?

Incorporating prior knowledge

More freedom could reduce error but could we manage this new complexity?

$$\hat{\beta}^{\text{prior}} = \text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \sum_{j=1}^{p} \lambda_j |\beta_j|^q \right\} \quad (16)$$

Cons Opinion of Rich Sutton (see his webpage “The Bitter Lesson” (2019)): *And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation.*

Bias variance trade off

$$\begin{aligned} E[L] &= \int (y(x) - E[y] + E[y] - E[t|x])^2 p(x)dx + \int (E[t|x] - t)^2 p(x)dx \\ &= \int (E[y] - E[t|x])^2 p(x)dx + \int (y(x) - E[y])^2 p(x)dx + \int (E[t|x] - t)^2 p(x)dx \\ &= \text{Bias}^2 + Var(y(x)) + \text{noise}. \end{aligned} \tag{17}$$

Bias variance trade off

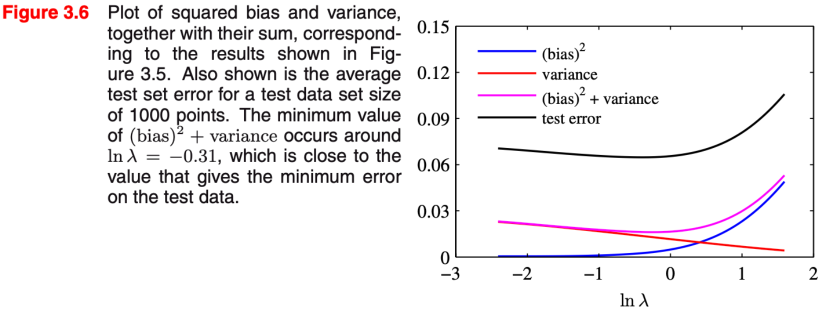
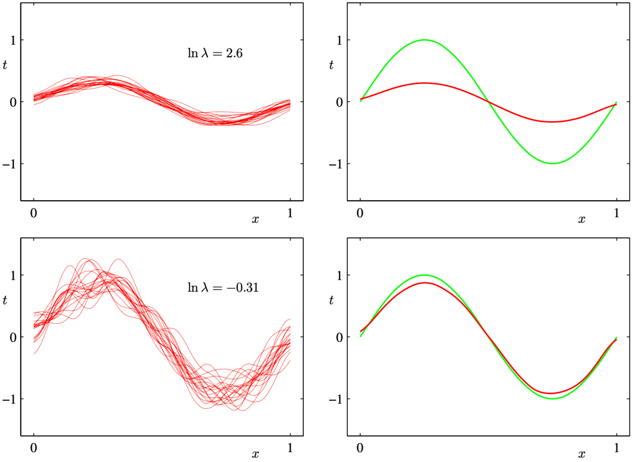


Figure 3.6 Plot of squared bias and variance, together with their sum, corresponding to the results shown in Figure 3.5. Also shown is the average test set error for a test data set size of 1000 points. The minimum value of $(\text{bias})^2 + \text{variance}$ occurs around $\ln\lambda = -0.31$, which is close to the value that gives the minimum error on the test data.

Figure 9: Error decomposition

How affect lambda



How affect lambda

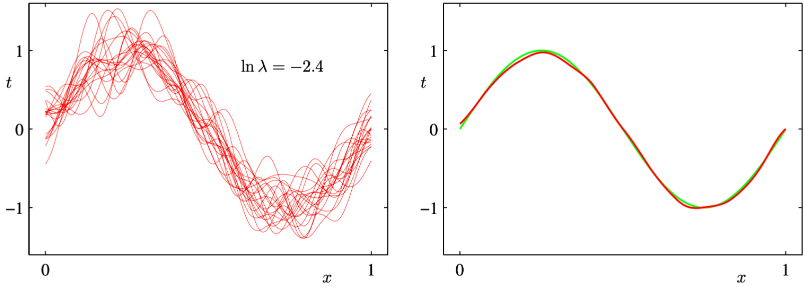


Figure 3.5 Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ, using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln \lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

What would we do next week

- Classification

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

“Matriz Definida Positiva.” 2022. 2022. https://es.wikipedia.org/wiki/Matriz_definida_positiva.

“Producto Escalar.” 2022. 2022. Producto escalar.

“Quadratic Programming.” 2022. 2022. https://en.wikipedia.org/wiki/Quadratic_programming.

“Quantile Regression.” 2022. 2022. https://en.wikipedia.org/wiki/Quantile_regression.

“Quantile Regression in Machine Learning.” 2018. 2018. https://towardsdatascience.com/quantile-regression-from-linear-models-to-trees-to-deep-learning-af3738b527c3.

Rodrigues, Filipe, and Francisco C. Pereira. 2018. “Beyond Expectation: Deep Joint Mean and Quantile Regression for Spatio-Temporal Problems.” arXiv. https://doi.org/10.48550/ARXIV.1808.08798.

“The Bitter Lesson.” 2019. 2019. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.