Predicting ozone levels

Simon Janežič

Introduction

Our goal was to build a Bayesian model for predicting ozone levels of seven weather stations located in Slovenia at 8 o'clock. We have about five years worth of data for all stations. Each data point includes many relevant features like past ozone measurements, meteorological measurements (e.g. temperature, wind) and meteorological predictions from an existing model (ECMWF).

Data preprocessing

We have removed date from features and replaced day of the year feature with cosine transformation of it that indicates seasonality.

$$seasonal\_feature = \cos\left(\frac{(day\_of\_year-1)2\pi}{364} - \frac{2\pi}{12}\right)$$

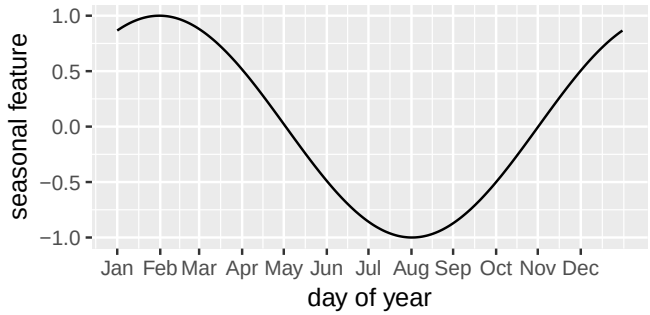

Figure 1: Day of the year feature to seasonal feature transformation. Ticks on x-axis indicate beginning of the months.

In Figure 1 we can see that the coldest months (January and Febraury) have transformed values of around 1, while the hottest months (July and August) have transformed values of around -1.

As final data preprocessing steps we removed features with zero variance and standardized features that were left. We also removed samples with unknown attribute values.

Model comparison

We have tested four different linear bayesian models:

- **linear regression (LR)** - basic linear regression (one model per station).
- **lasso regression (L1)** - L1 regularized linear regression with a hyperprior on regularization parameter (one model per station).
- **gamma regression (GR)** - L1 regularized gamma regression with a log link function and hyperprior on regularization parameter. The motivation for using gamma regression is the fact that the domain of our target variable is positive (one model per station).
- **joint lasso regression (JL1)** - L1 regularized linear regression with a hyperprior on regularization parameter. Uses data from all stations to predict ozone levels of every station. We have achieved that by simply adding a station identifier attribute, which is broken into 6 binary attributes (7 stations).

To evaluate our models we used data from the most recent year of every station as a test set. Log likelihood was used as a metric.

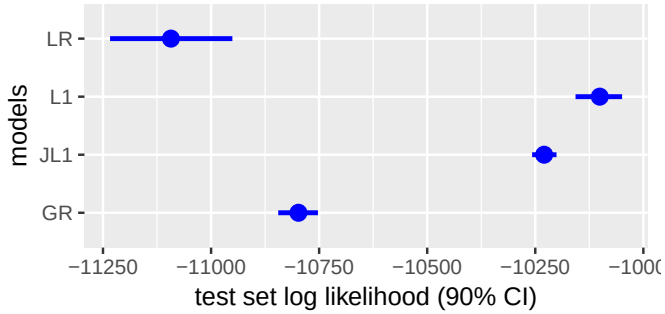

Figure 2: Model comparison. Dots represent expected values and lines represent 90% CI.

In Figure 2 we can see that LR seems to perform the worst. It might be overfitting since adding regularization (L1) gives us model with visually the best score. The fact that JL1 model has lower expected score than L1 means that the effects of meteorological attributes might not be independent from station's location and simply encoding station as a discrete attribute is not enough to make up for that. Since the two models are visually fairly close we also estimate the probability that L1 model has higher test set log likelihood: $P(L_t(L1) > L_t(JL1)) \approx 0.999 \pm 0.001$. This reaffirms our belief in L1 as the strongest model. Gamma regression's expected log likelihood is noticeably lower than the two best models, indicating that gamma distribution is not the right choice for predicting ozone levels.

Comparing performance on stations

For this part of the analysis, we take our best model (L1) and try to answer if and how it's performance changes from station to station. Log scores had to be normalized due to small differences in number of samples between the stations.

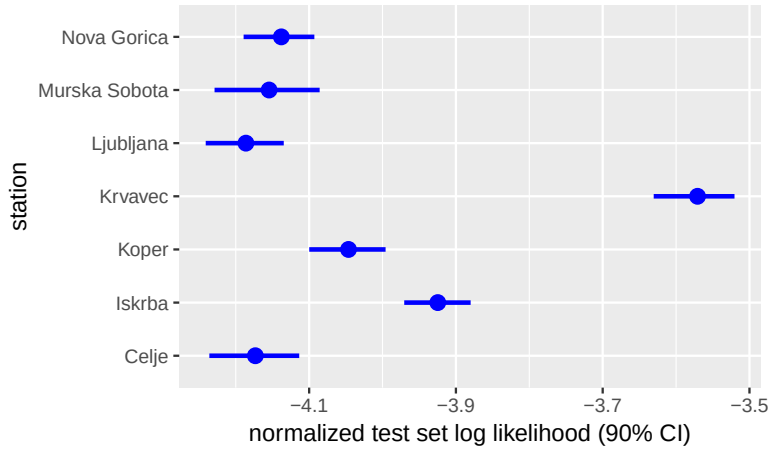

Figure 3: L1 performance on different stations. Dots represent expected values and lines represent 90% CI.

In Figure 3 we can spot some differences between the stations. Most noticeably Krvavec seems to be easiest to predict. Krvavec also has by far the highest altitude (1740m) and is not a city. Next outlier would be Iskrba, which also has relatively high altitude (540m) and is not a city. All the other stations are located in cities which have low altitude and most of them have comparable performance, indicating that ozone levels in cities are more unpredictable (with current features). One exception would be Koper that is somewhat easier to predict, but since this is a city by the sea it might be a special case.

Model explanation

In this section we try to find and explain the most important features when it comes to predicting ozone levels. We use our best model (L1).

Finding most important features reliably is a challenging task since a lot of the feature pairs are very highly correlated (Figure 4). However, we cannot ignore the fact that we used L1 regularization on our model. This means that highly correlated features are a lot less likely to obtain large weights that effectively cancel each other out for minor performance improvements, since our Laplace prior will outweigh that. That is, if our regularization parameter (Laplace distribution scale) is strict enough.

We can estimate regularization parameter's expected value for every station: $E[\lambda] \approx [0.70 \pm 0.00, 0.51 \pm 0.00, 0.51 \pm 0.00, 0.35 \pm 0.00, 0.66 \pm 0.00, 0.58 \pm 0.00, 0.71 \pm 0.00]$. All features are standardized, 90% sample CI for target variable is $(28.0, 142.6)$ and expected values for all lambda parameters are below 1. From this relatively large scale difference we can intuitively say that regularization is quite strict.

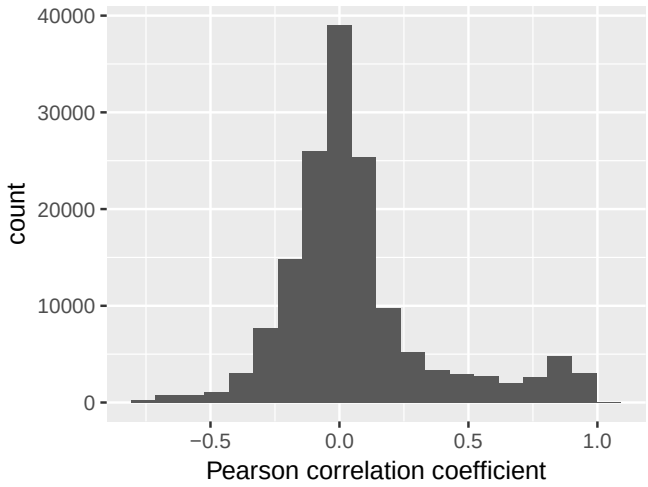

Figure 4: Attribute pairs Pearson correlation histogram. Correlations of variables to themselves are not included in the histogram.

To find most important features we calculated expected absolute weight values for every feature and sorted features in descending order. We used MCMC samples from all stations to achieve that, since we wish to find attributes that are important for all stations. Figure 5 displays the results. We notice that only a few of the attributes stand out with high expected importance. After those attributes the importance starts falling very slowly.

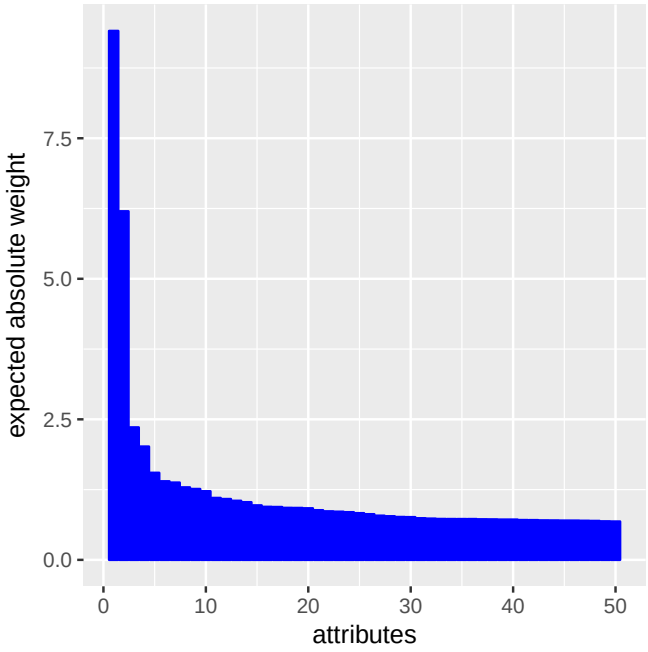

Figure 5: Expected importance of top 50 attributes.

Table 1: Six attributes with highest expected importance.

Attribute	**Expected absolute weight**
max. O_3 yesterday	9.41 ± 0.47
max. O_3 till 7h	6.20 ± 0.47
temp. difference (2m and 950 hPa) till 7h today	2.36 ± 0.21
max. global radiation till 7h	2.02 ± 0.13
max. temp. till 7h	1.55 ± 0.22
seasonal feature	1.40 ± 0.11

Those few important features that stand out are listed in Table 1. Two features with by far the highest expected importance have to do with ozone levels from the past. It is no surprise that ozone levels from the near past strongly affect future ozone levels. Perhaps it is more interesting that maximum values are deemed more important than the means. The importance of temperature features is no surprise either, since temperature and ozone are closely related. More ozone means more heat from the UV waves gets trapped in the stratosphere. At the same time ozone absorbs infrared radiation from the earth's surface which heats up the troposphere. Of course global radiation plays a big part in the very same process, as does the seasonality (more sun during summer, more pollution during winter).

Conclusion

We've shown that basic linear regression can be significantly improved with L1 regularization when predicting ozone levels from the given dataset. We also noticed that ozone levels are easier to predict for certain stations and found a potential link with the station altitude and a station being in a city or in a rural area. Finally, we found a reasonable explanations for the few most important attributes whose expected absolute weight values stood out.