Tasks chapter4: classification

David Kurz, Malte Hildebrandt
2023S707737 VU Geostatistik
Universität Innsbruck

April 18, 2023

Task 1: Classification: forecast of wet/dry mornings

This task will look at mornings (06 - 12 UTC) that had measurable precipitation > 0.1 mm.
Again, use the SYNOP observations and forecast data from ECMWF with a forecast horizon of 36 hours for the "location" assigned to your dyad, contained in location_obs_ECMWF_2009-2022.rds. In the following text, replace location with the name of the one assigned to you, e.g. ibk. The rds-file contains several forecast parameters for which acronyms, description and units are given in ECMWF_selected_parameters.pdf.

1. **Data preparation:** Add a column with a binary variable location$wet to your location zoo-object; it should be 1 for a wet morning (location$rr6hObs > 0.1) and 0 for a dry morning. Use ifelse(). You might want to remove NA values.

---

2. **Base rate:** What is the base rate of wet mornings?
"Base rate" is the probability of an event without using any additional information. In this case it is the fraction of 6-h mornings which had observed measureable precipitation. Exploiting that we gave a wet morning a value of 1 and a dry morning a value of 0, the number of wet mornings can be simply computed with sum(location$wet) and the total number of mornings with length(location$wet). Alternatively, sum(location$rr6hObs > 0) will also work for the number of wet mornings since a logical inquiry results in 1 or 0.

   $\rightarrow$ At the location of Linz Hoersching we got a base rate of 0.233 of wet mornings.

---

3. **Splitting data set into training and testing parts:** Subset your data to use even mornings (i.e. mornings of days 2, 4, 6, 8, . . .) for training and assign it to trainLocation, and use odd mornings (of days 1, 3, 5, 7, . . .) for testing (assign to locationTest). You can find even and odd days with the modulo function %% 2.

---

4. Classify: Logistic regression belongs to the family of generalized linear regression models, which can be fit in R with glm. To select logistic regression you have to use it with the option family = binomial. Fit a logistic regression with only 1 predictor variable, the ECMWF 6-h precip sum location$rr6h to the training data and assign the model to the variable glm.fit. Remember that specifying a model in R is simple with the tilde $\sim$ character: response $\sim$ predictor1 + predictor2 ...

---

5. **Goodness of fit:** summary() of the model fit on training data. The z-statistic is the equivalent of the t-statistic in linear regression. It is the coefficient divided by its standard error (cf. James2021, p 136). Note, that the (absolute value of the) z-statistic is the number of standard deviations away from the mean. An absolute value of at least 1.96 indicates a p-value of 0.05: only in 5% of the cases could the null hypothesis be true by chance. Is there an assocation between observed wet mornings and their forecast (i.e. can the null hypothesis be rejected)? Further explanation of the output of summary(glm.fit): deviance generalizes RSS (residual sum of squares) to a broader class of models (it is the maximized log-likelihood times -2). The smaller the deviance, the better the fit. The null deviance shows how well the response is predicted by the model with only an intercept. The residual deviance shows how well the response is predicted by the model when predictor(s) are included. Degrees of freedom: number of observations minus number of predictors.

Table 1: Deviances

	null deviance	**residual deviance**
train data	2463.4 on 2514 d.o.f.	1821.8 on 2513 d.o.f.
train multi data	2463.4 on 2514 d.o.f.	1715.1 on 2511 d.o.f.
test data	2463.4 on 2514 d.o.f.	1822.1 on 2513 d.o.f.
test multi data	2463.4 on 2514 d.o.f.	1769.1 on 2511 d.o.f.

$\rightarrow$ The null-deviance considers the model prediction of data without any predictors. In that case it is the same value for all data sets, calculated from the average frequency of an event. Here it is a high value, so the model prediction is very bad without considering predictors. The residual-deviance consideres model predictors. Therefore the model prediction is better the more predictors are used. Training a model also improves its prediction performance.

Table 2: Coefficients

	Coefficient	**Std. error**	**z-statstic**	**p-value**
	train data			
Intercept	-2.28253	0.07395	-30.87	<2e-16
tp6h	1255.64158	71.57569	17.54	<2e-16
	train multi data			
Intercept	-2.7095	0.1267	-21.382	<2e-16
tp6h	497.1767	93.3297	5.327	9.98e-08
lsp6h	1698.8894	187.9496	9.039	<2e-16
tcc	0.6865	0.1670	4.110	3.95e-05
	test data			
Intercept	-2.28916	0.07442	-30.76	<2e-16
tp6h	1400.64128	81.08613	17.27	<2e-16
	test multi data			
Intercept	-2.6270	0.1190	-22.075	<2e-16
tp6h	781.8172	109.0415	7.170	7.51e-13
lsp6h	1279.9462	198.3642	6.453	1.10e-10
tcc	0.6032	0.1608	3.752	0.000175

→ The absolute value of z-statistic is higher for training data with only one predictor. If we introdruce more preditors and test the models data we get lower absolute values from it. High z-value notes that observed values are further away from the median and it is unlikely having a systematic failure.

→ Due to a very small p-value, the null-hypothesis has to be rejected. Only in a very view cases, maximum of 0.0175 %, the null-hypothesis could be true by chance. So there should be an association between observed wet mornings and their forecast.

Table 3: Deviance Residuals

Min	**1Q**	**Median**	**3Q**	**Max**
train data				
-5.2806	-0.4614	-0.4408	-0.4408	2.1816
train multi data				
-5.4045	-0.4964	-0.4169	-0.3590	2.3554
test data				
-5.7141	-0.4594	-0.4394	-0.4394	2.1882
test multi data				
-6.8161	-0.4975	-0.4180	-0.3736	2.3245

→ The values of table 3 show that the median is lower with more predictors for training data as well as for testing data. The more predictors we have, the further apart are Minimum, 1st and 3rd quantile and Maximum. The training data has a slightly better deviance than the test data.

6. **Predict:** Use the model glm.fit to make predictions on the training data set (assign to predictTrain) and test data set (predictTest), respectively using predict(), which takes as arguments the model (glm.fit, the data (trainLocation or testLocation) and type = "response".

---

7. **Confusion matrix:** Compute the confusion matrix separately for predictions on the training data set and the test data set, respectively, using a probability p=0.5 as threshold. What does the confusion matrix tell you about the types of mistakes that the logistic regression model makes in this case? Use table() to compute the confusion matrix. The option deparse.level = 2 will provide a complete labeling of the confusion matrix and its columns.

Table 4: Confusion Matrix

predicted	**truth**	**test**	**train**	**test multi**	**train multi**	
dry	dry	1975	1978	1982	1994	true negative
wet	dry	56	53	49	37	false negative
dry	wet	305	294	298	274	false positive
wet	wet	179	190	186	210	true positive

→ True negatives are the number of correctly predicted dry instance, whereas true positives are the number of correctly predicted wet instances. False negatives are the number of wet instances that were incorrectly predicted as dry. Finally the false positives are the number of dry instances that were incorrectly predicted as wet.

→ In general the confusion matrix allows to see the model's strengths and weaknesses. It can help identifying areas where the model needs improvement. If there is a large number of false negatives in the test set, it suggests that the model has difficulty correctly predicting wet instances. On the other hand, a large number of false positives suggests that the model has difficulty predicting dry instances accurately.

---

8. **ROC curve:** Compute the ROC for training and test data and plot both lines in one figure (cf. James2021, Fig. 4.8). Use the true positive rate on the y-axis and the false positive rate on the x-axis. See tables 4.6 and 4.7 in James2021 and - more exhaustive - the Wikipedia entry for confusion matrix for all the different metrics (and their multiple names) that can be computed from the confusion matrix (also known as "contingency table"). For example, true positive rate (== true positives divided by observed positives) is also called sensitivity, and true negative rate (== true negatives divided by observed negatives) is also called specificity. false_positive_rate is 1 - specificity. What are the values for the area under the curve (AUC)? How much worse is the performance on the test data compared to the training data? Think of an end-user for whom a high true positive rate is crucial, who is willing to tolerate a false positive rate of 20%, and who refuses

to get a probability but wants a yes/no of "wet" forecast. What would be the optimal probability to convert the probabilistic forecast into a binary one for this end-user? Many packages exist in R to plot ROC-curves. Use the package ROCit to compute the ROC curve. Comment on the width of the confidence interval. The vignette of the package gives a good overview of how to use it: https://cran.r-project.org/web/packages/ROCit/vignettes/my-vignette.html. First fit a ROC-object with rocEmp <- rocit() using the empirical method. class are the data you used for classifying the data, i.e. the observed wet vs. dry mornings; score is your predicted model (here: predictions on the training data and test data, respectively). Since rocit cannot handle the time information contained in the zoo-object, hand over just the data without the date information with coredata, e.g. coredata(predictTrain). Also suppress plotting the legend with the option legend = FALSE. Use names(rocTest) to see which variables the roc-object contains. Then plot the ROC curve for the test data with lines(rocTest$TPR ~ rocTest$FPR, col = "red"). Notice that ~ indicates that you are plotting a function when you use this command! Finally add a legend with legend() and the options horiz = TRUE, lwd = 4 and the AUC-values (with text() and paste() to concatenate "AUC:" and the AUC value rounded to 3 digits: round(rocTrain$AUC,3)).

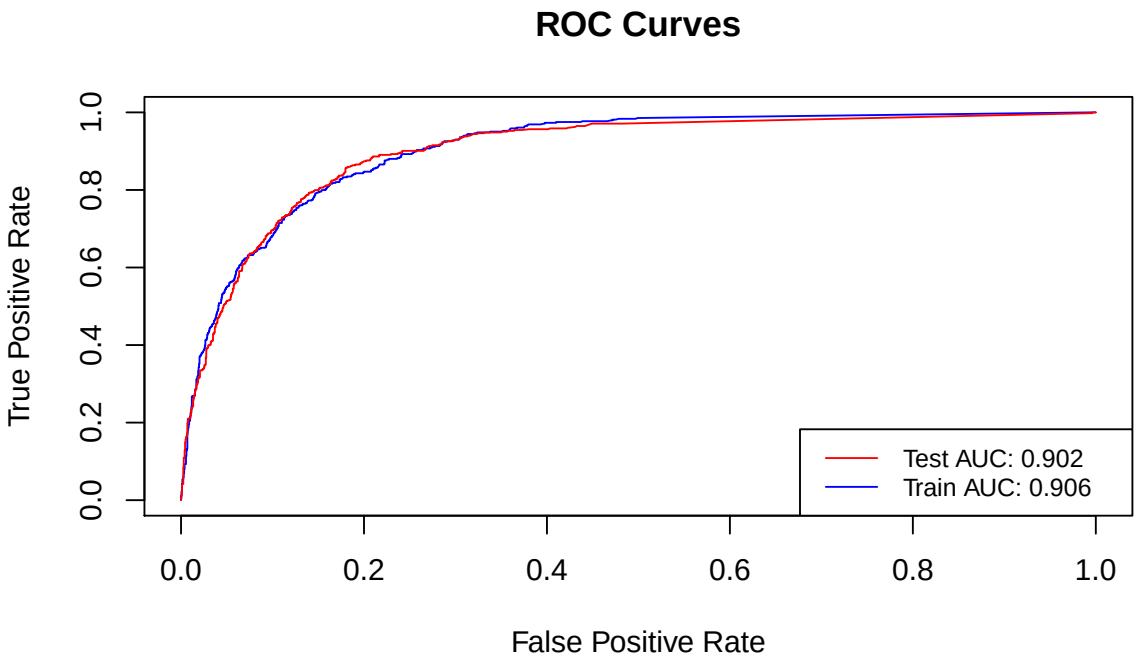

Figure 1

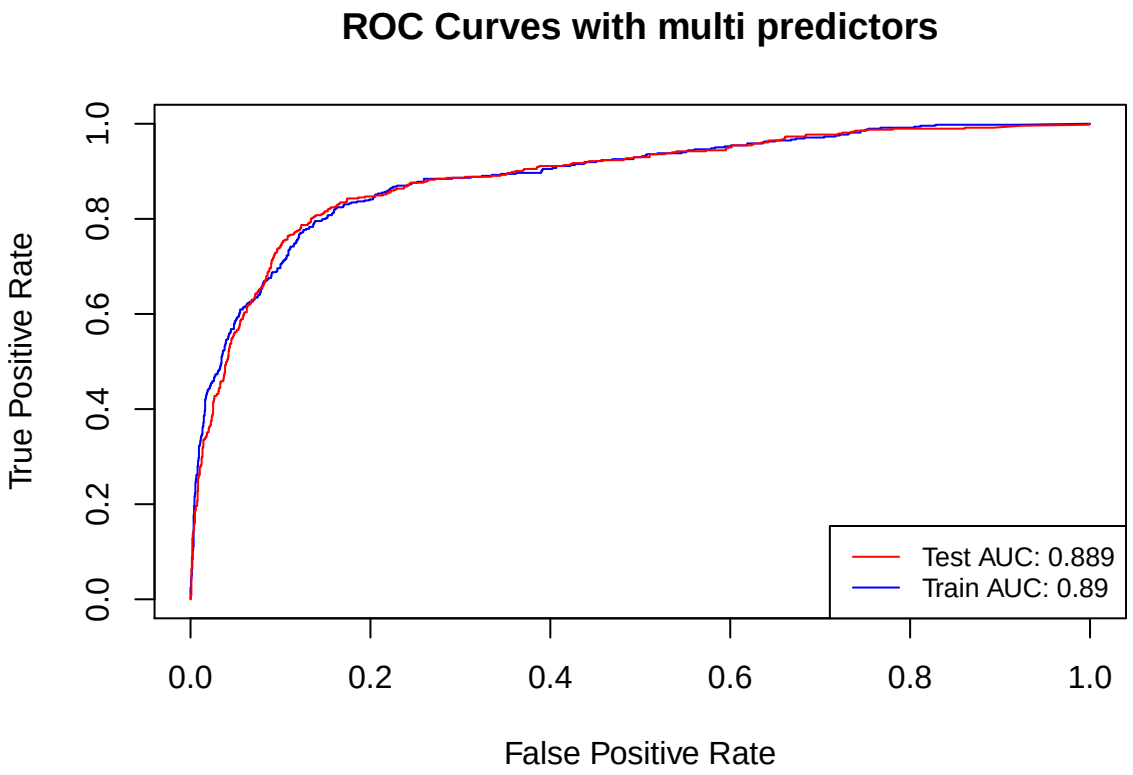

Figure 2

→ How much worse is the performance on the test data compared to the training data?
The training data performce slightly better than the test data (0.1%).

→ Think of an end-user for whom a high true positive rate is crucial, who is willing to tolerate a false positive rate of 20%, and who refuses to get a probability but wants a yes/no of "wet" forecast.
(high true pos.) harvesting, **(tolerate)** snow clearance service, **(yes/no)** outdoor events

→ What would be the optimal probability to convert the probabilistic forecast into a binary one for this end-user?
50 % as threshold point (as done in WB). Eather we take a probability above that or below that threshold. 50 / 50 is not welcome.

9. **Precision-recall (PR) curve:** Other information from the confusion matrix can be extracted and plotted, too. If the base rates of the 2 classes (wet and dry in our case) would differ strongly (say a hundred times more events in one class than the other), a precision-recall curve is better suited to display the skill of the model. “Precision” (or “positive prediction value”) is the number of true positives normalized by the sum of true positives and false positives. “Recall” is another expression for “sensitivity” or “true positive rate”. Plot a figure with precision as a function of recall for the model predictions for training and test data, respectively. Use type = "p", pch = 19, cex = 0.15 as plot options and add the test data PR curve with points().

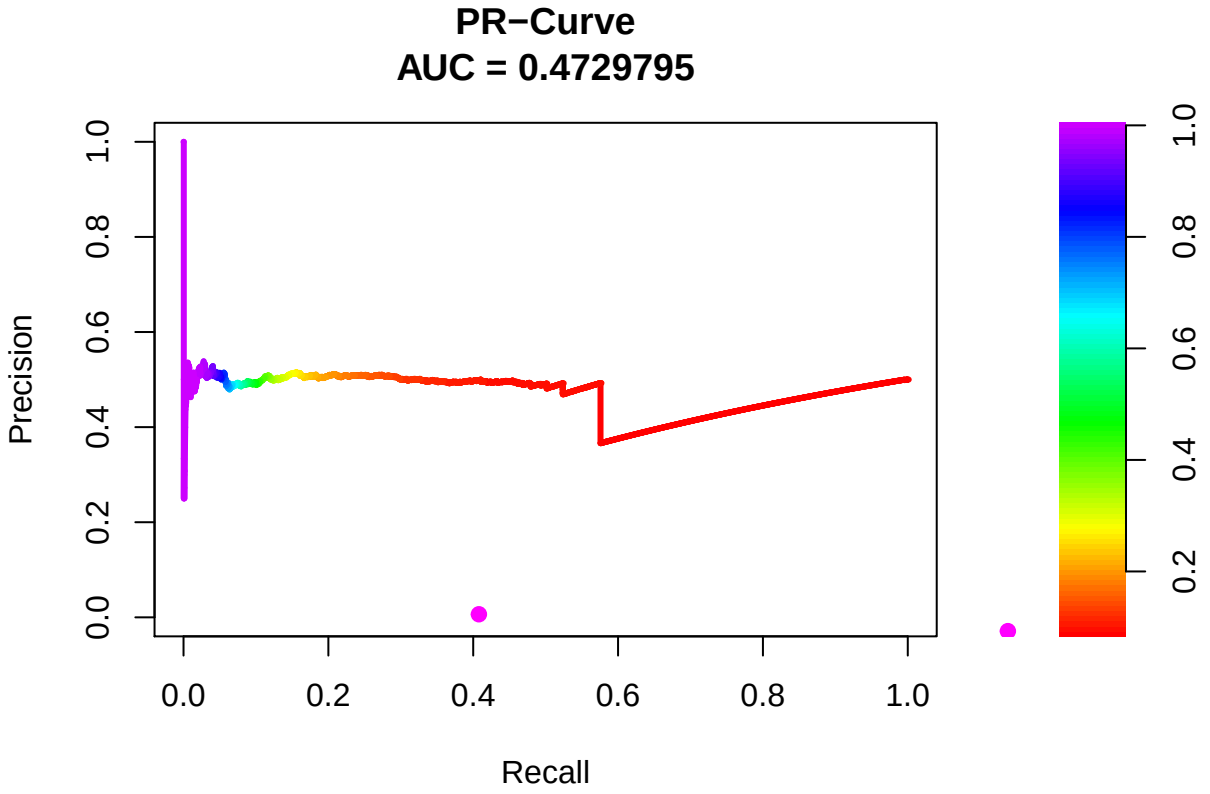

Figure 3

→ When base rates of the 2 classes (wet & dry) in a model are imbalaced the classic accuracy amtric can be misleading. In that case the classifier prediction will output a high accuracy. A perfect classifier would have a precision and recall of 1.0. Our plot depicts a bad classifier. With an AUC-value of close to 0.5 it seems having a random classifier. We understand this like every outcome predicted correct, the next prediction is incorrect and this not or just slightly associated with the outcome. In other words it’s very likely that a wet morning is followed by a dry one.

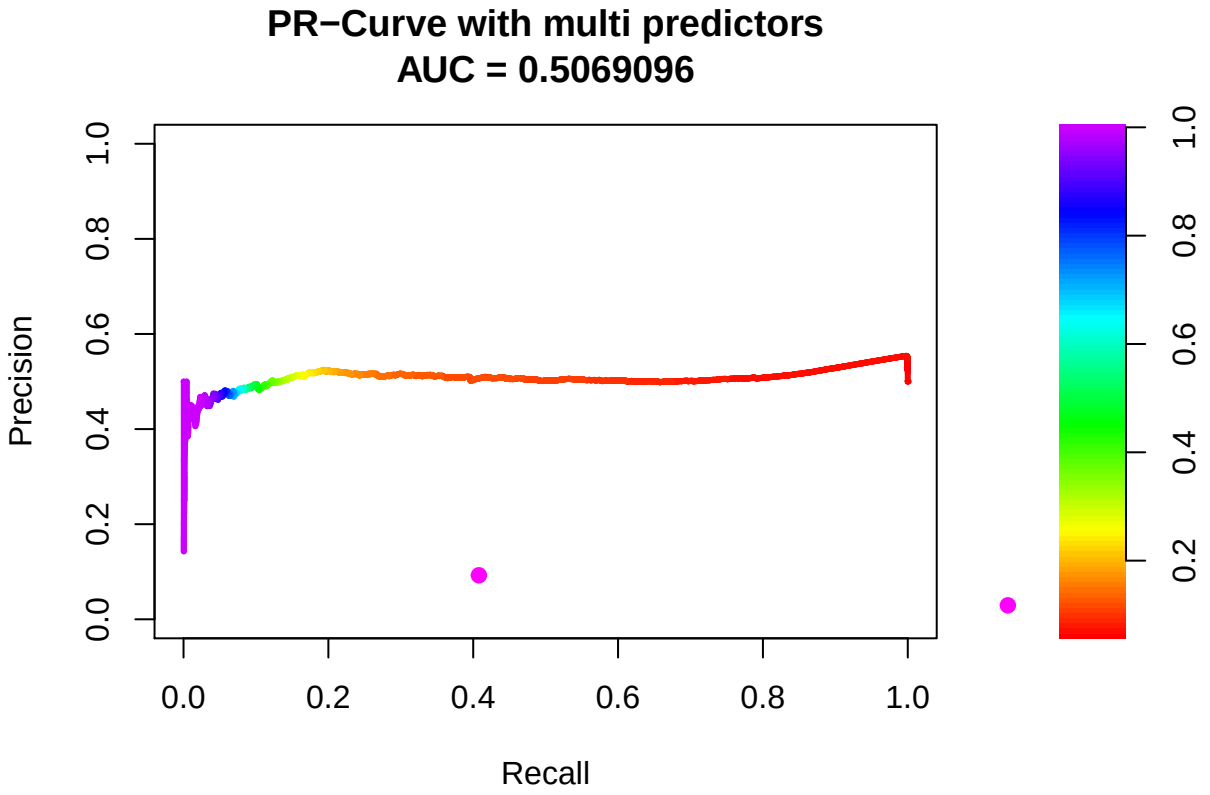

Figure 4

Task 2: Atmospheric/cryospheric examples for using Poisson regression

Think of an example from atmospheric or cryospheric science with count data (e.g. number of people using a bike offered by a bike sharing company as described in section 4.6.2 of James2021). What is the response and what are potential predictors? If you are from another Master's program, use an example from your field.

Example: CO2 content near the ground

→ response: concentration of CO2
→ predictors: f.e. day of the week, season, temperature, wind (direction, speed), precipitation, biosphere and solar radiation