Generalized Linear Regression and Classification

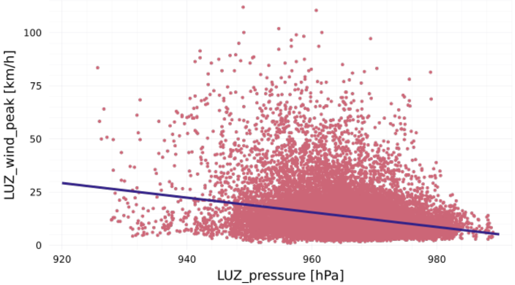
Johanni Brea

Einführung Machine Learning

GYMINF 2022

Table of Contents

Wind Speed Prediction



$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 346 \text{ km/h}, \theta_1 = -0.344 \text{ (km/h)/hPa}$$

- **Training Set**: Hourly data 2015-2018
- **Training Loss (rmse)**: 10.0 km/h
- **Test Set**: Hourly data 2019-2020
- **Test Loss (rmse)**: 11.5 km/h

root-mean-squared error:

$$\text{rmse} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Multiple Linear Regression

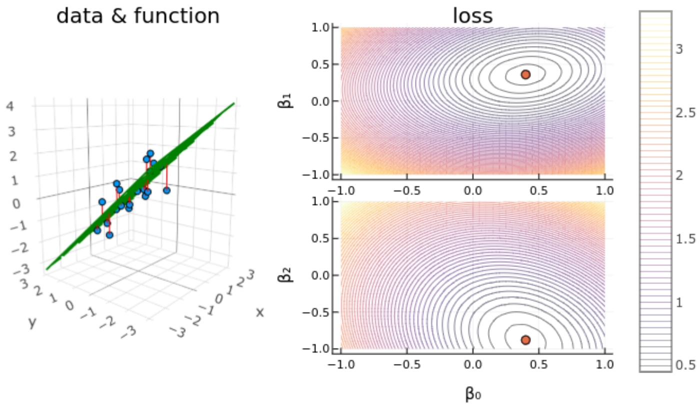
$$\hat{y} = f(x) = f(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}}_{\boldsymbol{X}} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Often the output correlates with multiple factors.

For example:
x_1: pressure in Luzern
x_2: temperature in Luzern
x_3: pressure in Basel
x_4: pressure in Lugano
etc.

Multiple Linear Regression Example: $p = 2, n = 20$



Multiple Linear Regression finds the plane closest to the data.

Multiple Linear Regression for Wind Speed Prediction

predictor name	fitted parameter
LUZ_pressure	-2.79 (km/h)/hPa
PUY_pressure	-2.39 (km/h)/hPa
BAS_precipitation	-0.66 (km/(h)/mm
⋮	⋮
LUZ_temperature	0.87 (km/h)/C
GVE_pressure	3.95 (km/h)/hPa

Interpretation

An increase of one hPa of LUZ_pressure correlates with a decrease of the expected wind speed by 2.79 km/h, if all other measurements remain the same.

Evaluation

- **Training Set**: Hourly data 2015-2018
- **Training Loss (rmse)**: 8.1 km/h
- **Test Set**: Hourly data 2019-2020
- **Test Loss (rmse)**: 8.9 km/h

Table of Contents

Error Decomposition for Regression

Conditional Data Generating Process: $Y = f(X) + \epsilon$

noise (or error term) ϵ, with expectation $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$
f represents **systematic** information that X provides about Y.

E.g. $f(X) = \sin(2X) + 2(X - 0.5)^3 - 0.5X$

$x_1 \approx 0.2 \quad f(0.2) \approx 0.23 \quad \epsilon_1 \approx -0.03 \quad y_1 \approx 0.2$

Supervised Learning: $\hat{Y} = \hat{f}(X)$

$\hat{f}$ = estimate of f, $\hat{Y}$ = predicted outcome

E.g. $\hat{f}(X) = 0.1 + X \quad \hat{y}_1 = \hat{f}(x_1) = 0.3$

residual $y_1 - \hat{y}_1$

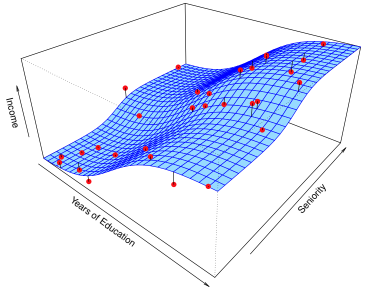
prediction error $(y_1 - \hat{y}_1)^2 = 0.1^2 = 0.01$

Blackboard: Error Decomposition for Regression

$$\mathrm{E}_{Y|X}(Y - \hat{Y})^2 = \underbrace{\left(f(X) - \hat{f}(X)\right)^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}$$

Quiz

- Let us assume the red data points in this figure were generated with the help of a function whose graph is shown in blue. What is correct?
  A. A linear fit has a non-zero reducible error.
  B. The irreducible error is zero.
  C. A method with zero prediction error on the red data has a reducible error of zero.



- Assume we perform linear regression on the red data points and compute the residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$ and the empirical variance $\frac{1}{n-1}\sum_{i=1}^{n}\hat{\epsilon}^2$. The irreducible error is
  A. larger B. smaller
  than this empirical variance.

- What is the irreducible error for the following data generating process?

$$p(y|x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-2x)^2}{2}}$$

A. 1 B. 2 C. 4

Summary

- The true systematic information f that X provides about Y is usually unknown.
- Our goal: find the function $\hat{f}$ that minimizes the reducible error.
- The test error of $\hat{f}$ is never lower than the irreducible error.

Table of Contents

Spam Classification

spam

Subject: follow up
here ' s a question i ' ve been wanting to ask you , are you feeling down but too embarrassed to go to the doc to get your m / ed ' s ?
here ' s the answer , forget about your local p harm . acy and the long waits , visits and embarassments . . do it all in the privacy of your own home , right now . http : / / chopin . manilamana . com / p / test / duet it ' s simply the best and most private way to obtain the stuff you need without all the red tape .

Feature Representation

There are many ways to extract useful features from text. Here we use a very simple "bag of words" approach: word counts for a lexicon of size p.

E.g.

X_1 (your)	X_2 (need)	X_3 (pay)	$\cdots$	X_p (red)
3	1	0	$\cdots$	1

All n emails get such a representation.

Multiple Logistic Regression

$$\Pr(Y = \text{spam}|X) = \sigma(\theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad \sigma(0) = 0.5 \quad \sigma(-\infty) = 0 \quad \sigma(\infty) = 1$$

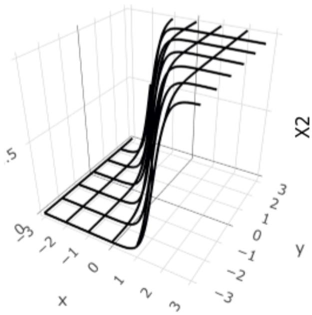
Find $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p$ that maximize the likelihood function.

Predictions (at **decision threshold** 0.5):
A new email is classified as spam, if its feature representation x leads to
$$\sigma(\hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_d x_d) \geq 0.5.$$

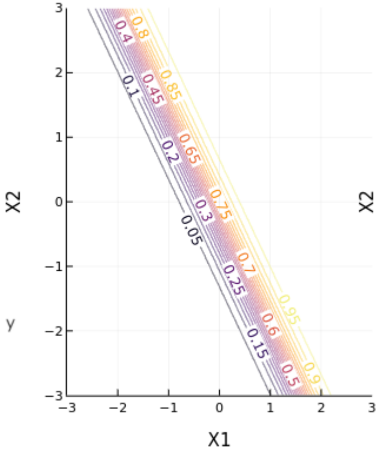
The corresponding **decision boundary** is linear:
$$\hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_d x_d = 0$$

Multiple Logistic Regression Example: p = 2

$\Pr(Y = \mathrm{A}|X)$ as 3D plot



$\Pr(Y = \mathrm{A}|X)$ as contour plot



samples and predictions

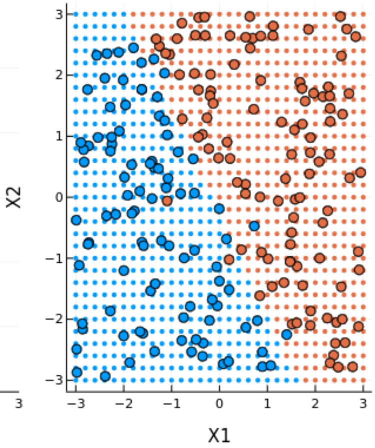
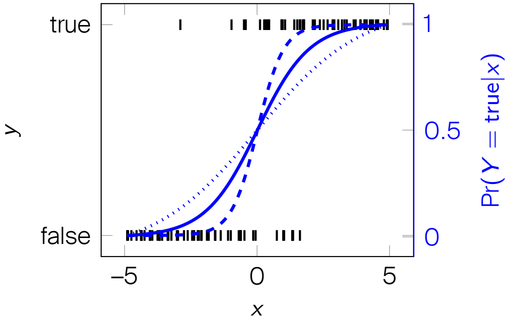


Table of Contents

Confusion Matrix



$\mathrm{Pr}(Y = \mathrm{true}|X = x) = \sigma(x)$ (solid)
$\mathrm{Pr}(Y = \mathrm{true}|X = x) = \sigma(2x)$ (dashed)
$\mathrm{Pr}(Y = \mathrm{true}|X = x) = \sigma(x/2)$ (dotted)

At decision threshold 0.5

		true class label		
		false	true	Total
predicted class label	false	42	4	46
	true	7	47	54
	Total	49	51	100

At decision threshold $\sigma(x) = 0.1$

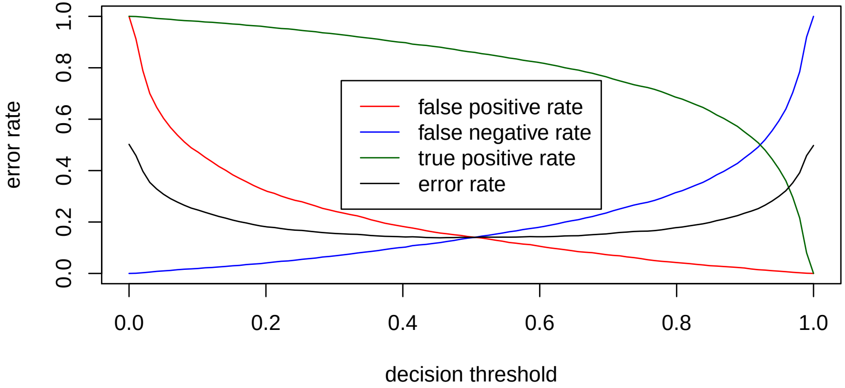
		true class label		
		false	true	Total
predicted class label	false	25	1	26
	true	24	50	74
	Total	49	51	100

Confusion Matrix & Error Rates

		true class label		
		Neg.	Pos.	Total
predicted class label	Neg.	True Neg. (TN)	False Neg. (FN)	N^*
	Pos.	False Pos. (FP)	True Pos. (TP)	P^*
	Total	N	P	

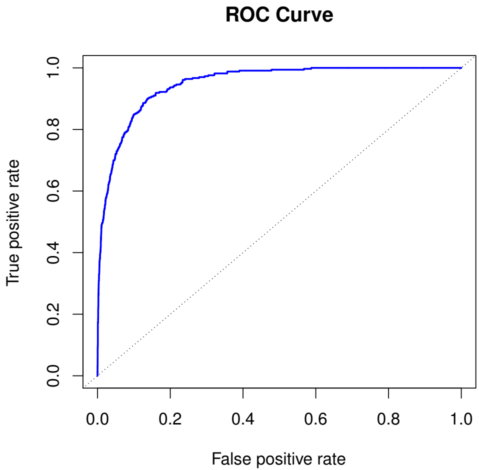
Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1-Specificity
True Pos. rate	TP/P	1-Type II error, Power, Sensitivity, Recall
False Neg. rate	FN/P	
Pos. Pred. value	TP/P^*	Precision, 1-false discovery, Proportion
Error Rate	$(FP+FN)/(P+N)$	Misclassification rate
Accuracy	1 - Error Rate	

Decision Thresholds and Error Rates



Finding the right threshold value depends on domain knowledge: which error do we most care about?
E.g. disease detection: do we want a small false negative rate?

ROC curve and AUC



- measure True Pos. rate and False Pos. rate for different thresholds on test data to obtain the receiver operating characteristics **ROC** curve.
- Random classification would be on diagonal.
- Area under the ROC curve **AUC** assesses the classifier.
- Random classifier has AUC = 0.5, perfect classifier has AUC = 1.

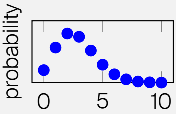
Quiz

1. Multiplying all parameters of logistic regression by a factor larger than 1 leaves the decision boundary unchanged.
2. If it is possible to perfectly classify the data, there exists a classifier with AUC = 1.
3. If we classify according to the worst classifier (class A if $p_A < 0.5$ and class B otherwise), the AUC is expected to be smaller than 0.5.
4. Typically we expect the AUC on the training set to be higher than on the test set.
5. No matter what classifier we use, the ROC curve always starts at (0, 0) and ends at (1, 1).

Table of Contents

Poisson Regression

Poisson



$$p(k|x) = \frac{e^{-f(x)} f(x)^k}{k!}$$

$f(x)$: a number

mean: $f(x)$

variance: $f(x)$

mode: $\lfloor f(x) \rfloor$ (floor)

When the response is a non-negative count variable, e.g. number of bicycles rented, it can be problematic to use the normal distribution to model the noise, because the support of the normal distribution is not restricted to positive numbers and the variance is independent of the mean.
The Poisson distribution can be better suited in this case (see bike sharing example in the notebook).

Take-home message

Always ask yourself: which distribution is best to model the noise.