Reaktor

Predictive analytics: hands-on session

Dev day 4.10.2013
Helsinki, Finland

What is predictive analytics?

- **Analytics**
  - Actionable information from observations
  - Summaries, causality, predictions, graphs, ...
- **Predictive Analytics**
  - Predict future, unseen, or results of actions
  - *Models*, often probabilistic, are fitted to data
  - Statistics, machine learning, data mining

Predictive analytics in practice

- Problem definition with business understanding
- Data collection
- *Data* preparation and *exploration*
- *Modeling and evaluation*
- Deployment
- Real world applications:
  - Customer relationship management: Netflix
  - Health care: Patients at risk of getting disease
  - Finance: Risk and fraud detection

The Wine data

- Chemical measurements of variants of the Portuguese "Vinho Verde" wine
- For 1599 red and 4898 white wines
- **Objective**: Predict the color of wine based on the chemical measurements (see right panel)
- Notice: No data about grapes, brand, selling price etc. available

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

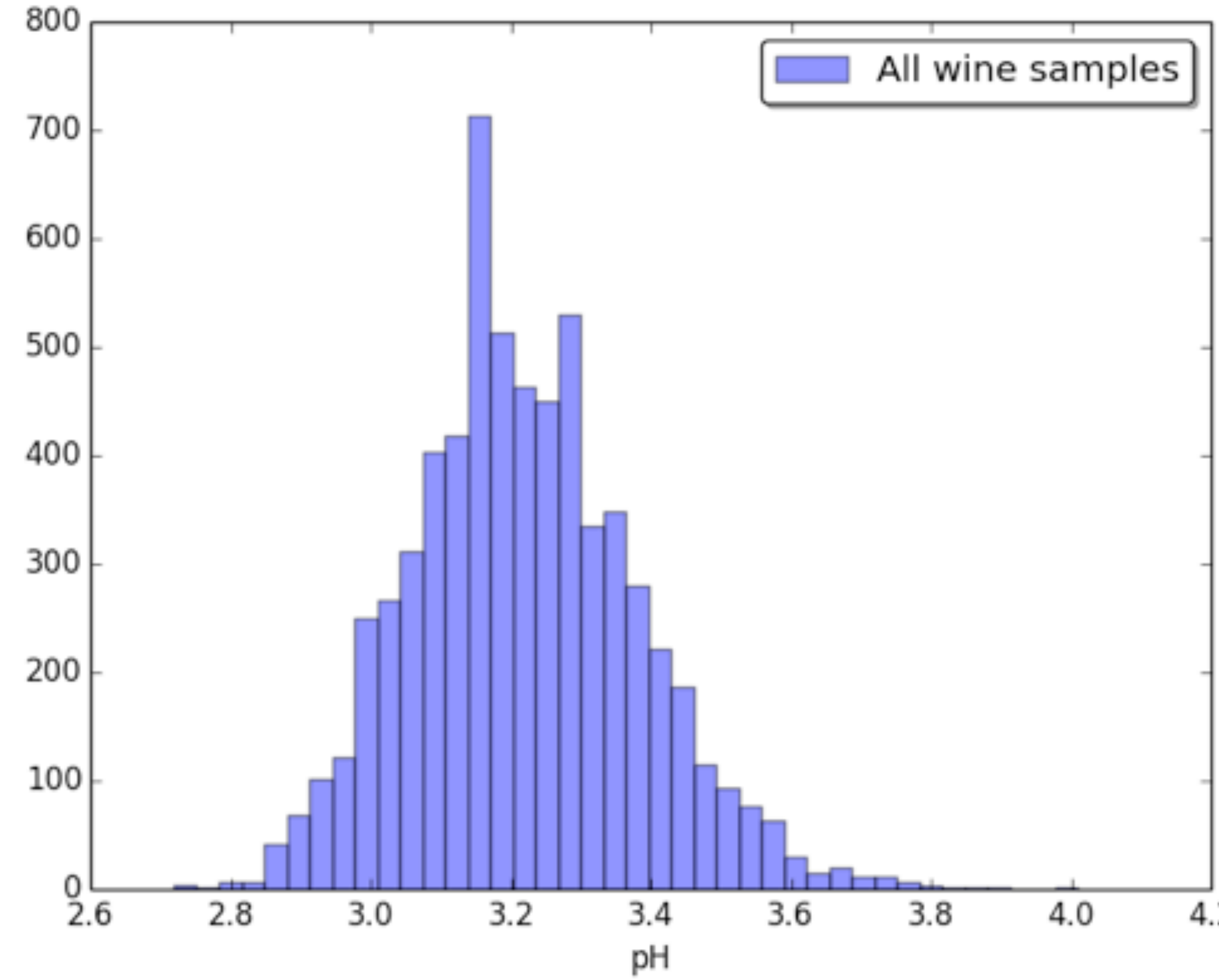
Reaktor

Know your problem and data

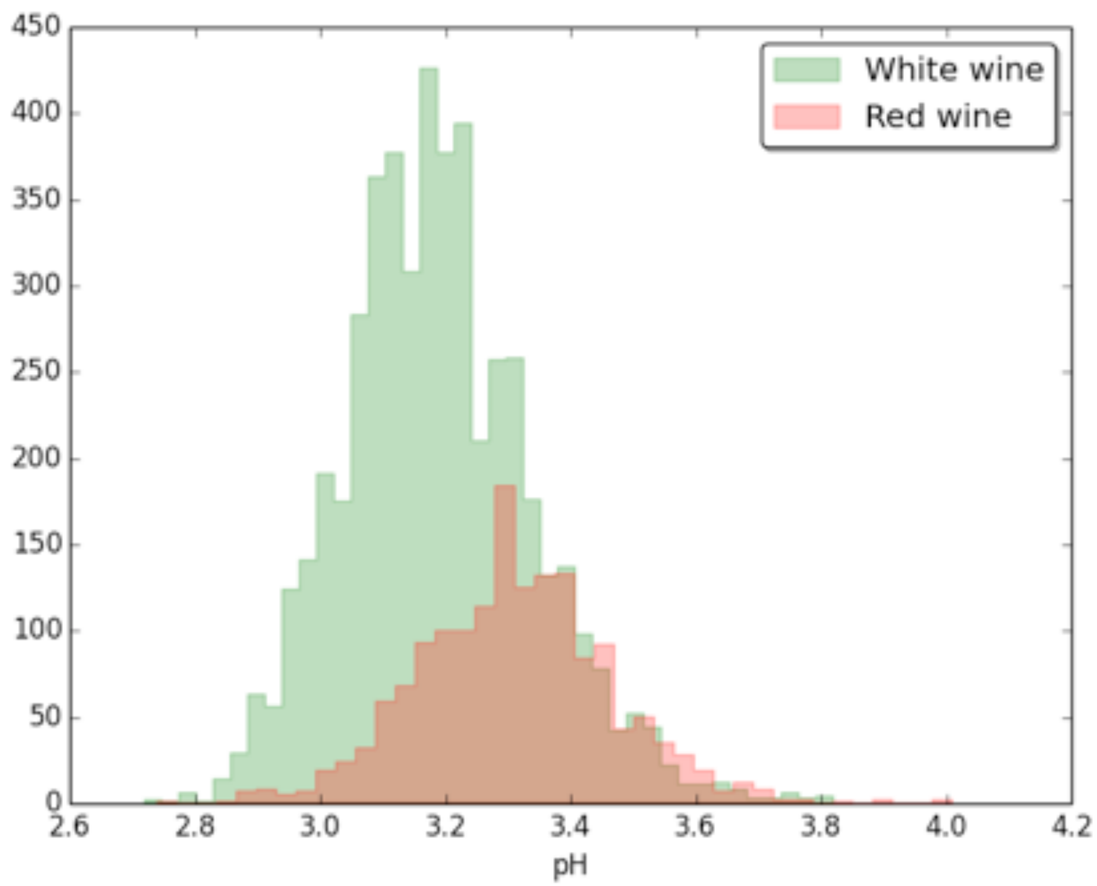
- Modeling requires context knowledge
  - Understand the domain
  - Get familiar with data
    - identify problems with quality
    - discover possibilities and limitations
    - *summary statistics and visualization*
  - Iterate over various models

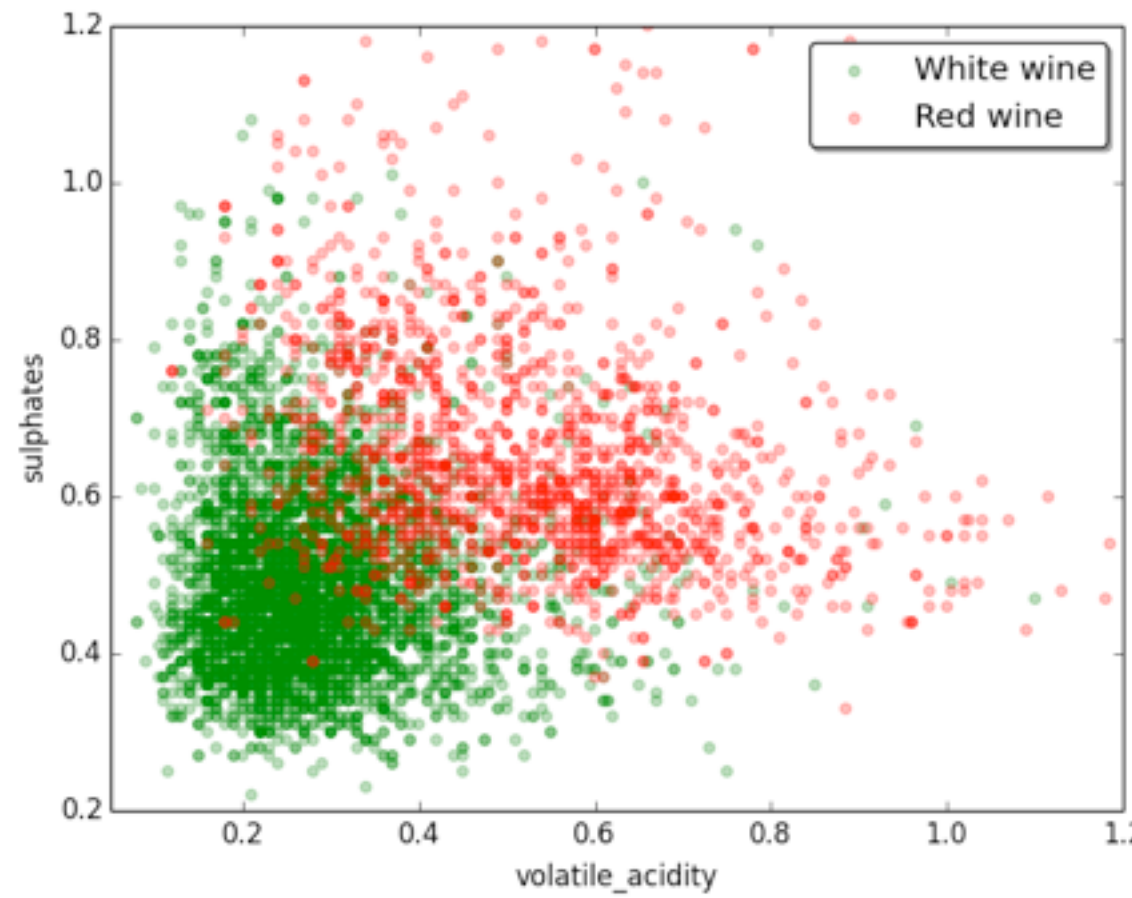
Insight to the data

- To make models, you need to know your data.
- Summary statistics
  - Mean (average)
  - Standard deviation
  - Quantiles
- Histogram
  - Estimate of the probability distribution
  - pH outside the range 2.9–3.6 is uncommon



Insights continued





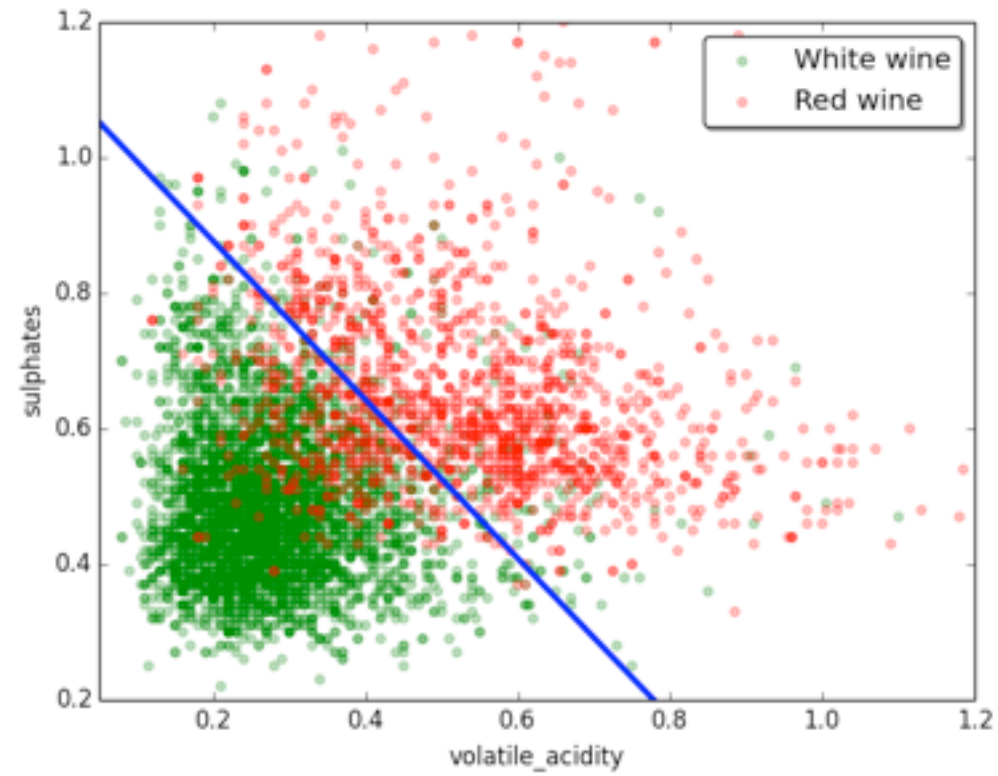
- Left: Histograms separately for red and white wines (1D)
- Right: A scatter plot
  - Not only variation but also covariation of two variables (2D)
- How about visualization in higher dimensions?

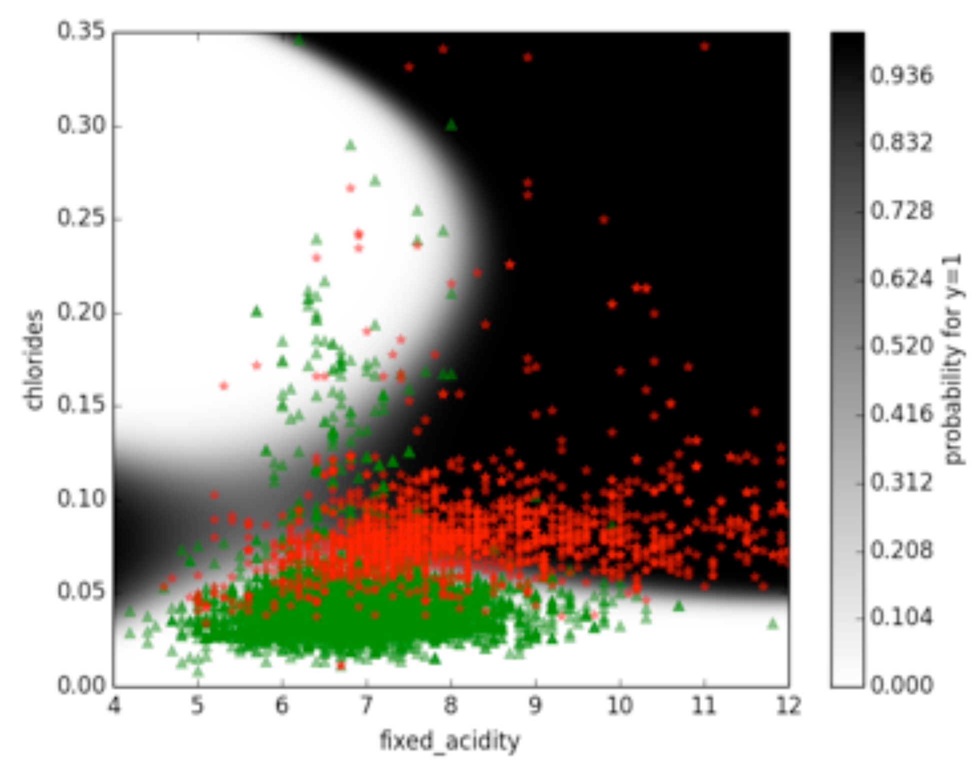
Hands on

- **The first session: Data exploration**
  - Read data from the file
  - Plot histograms of all variables
  - Plot histograms by wine type
  - Select two variables
  - Create a scatter plot using selected variables
  - See the script `hands-on.py`

Modeling: Logistic regression

- **Objective**: Find decision boundary to classify new samples
- To create a model:
  - Select a model class
  - Fit the model: Maximize the probability of the observed data (maximum likelihood)
- **Output**:
  - Probability of measurements representing a red wine
  - If probability > 0.5, the sample is classified as red wine





Reaktor

Logistic regression

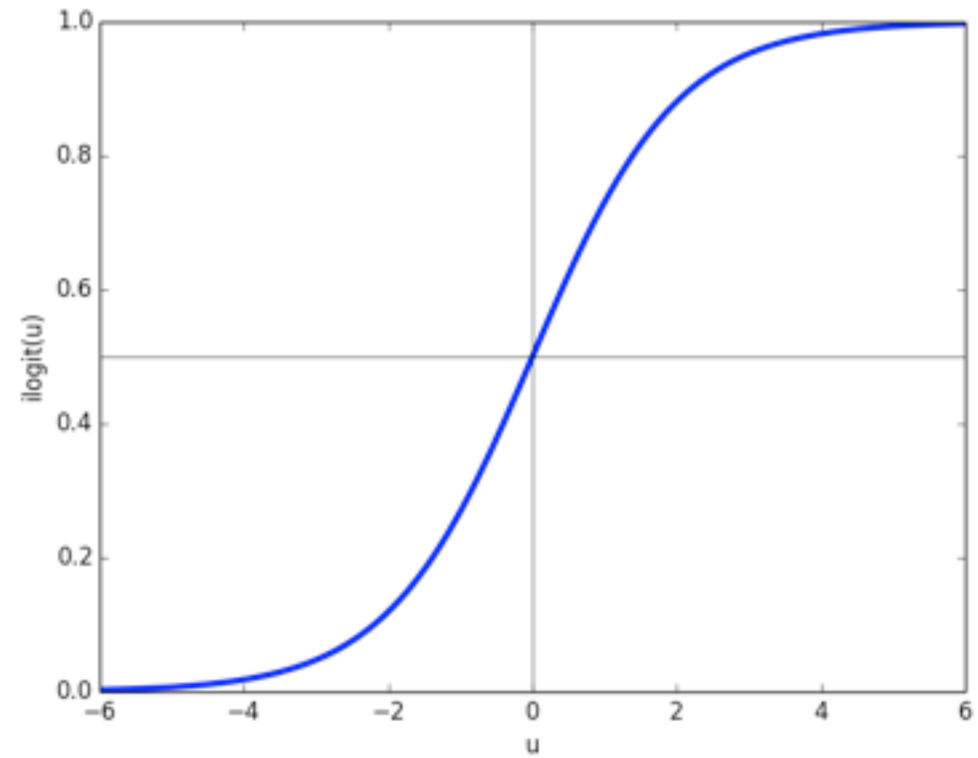
The conditional probabilities depend on *the linear combination of variables*. By inverse logit (ilogit) the outputs are *limited to the range [0,1]*.

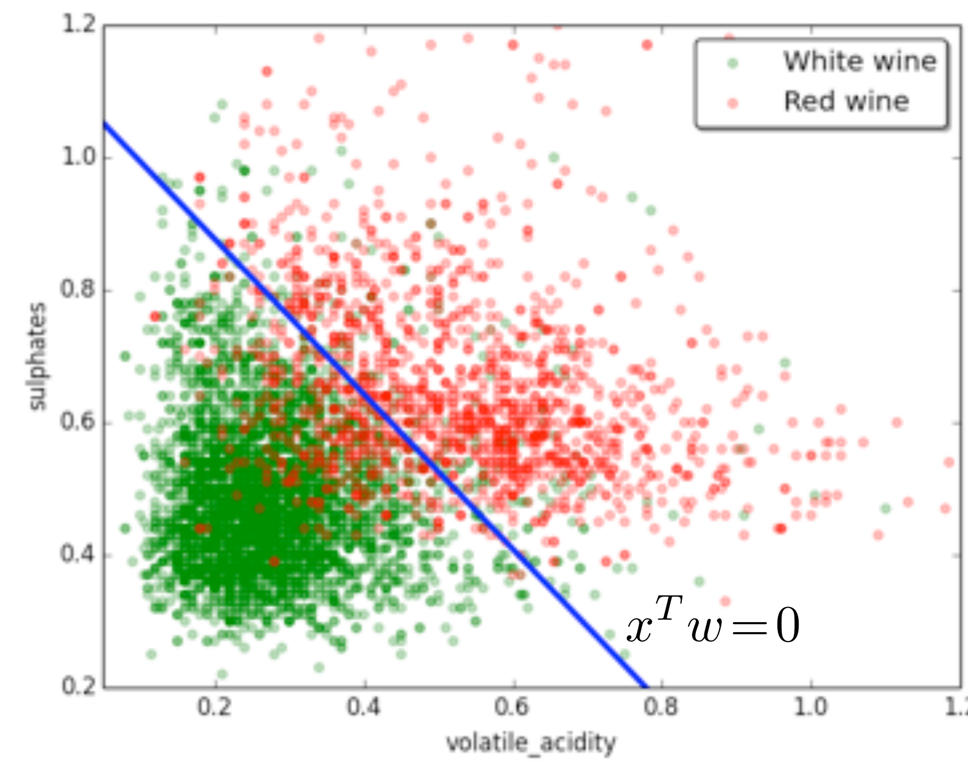
$$p(y=1|x,w) = \mathrm{ilogit}(x^T w)$$

where $$\mathrm{ilogit}(u) = \frac{1}{1+e^{-u}}$$

The model parameters are optimized by *maximizing the probability of observed data*. The maximum likelihood cost function is

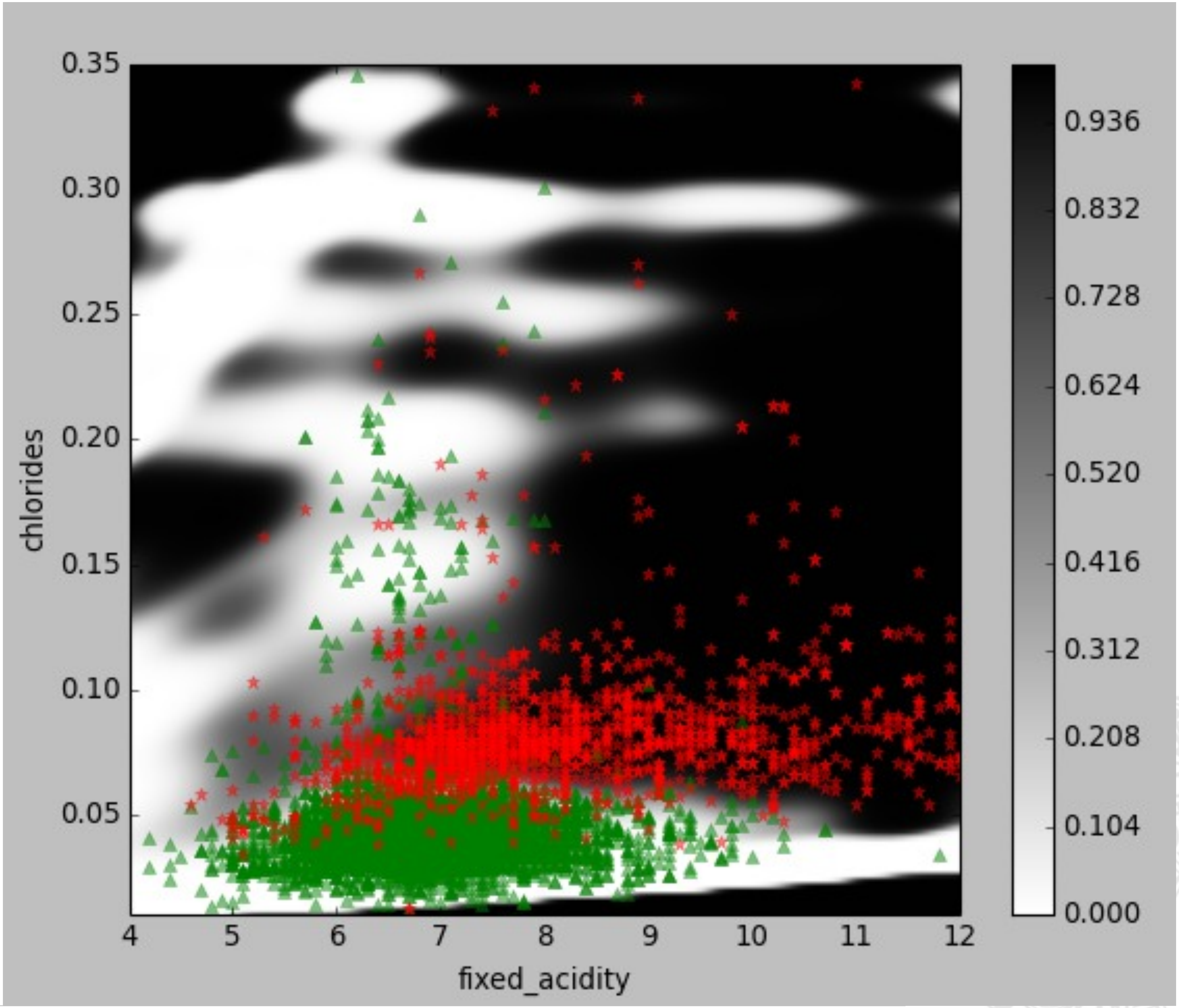
$$\max_w \prod_{i=1}^{N} p_i^{y_i} (1-p_i)^{1-y_i}$$





Hands-on continued

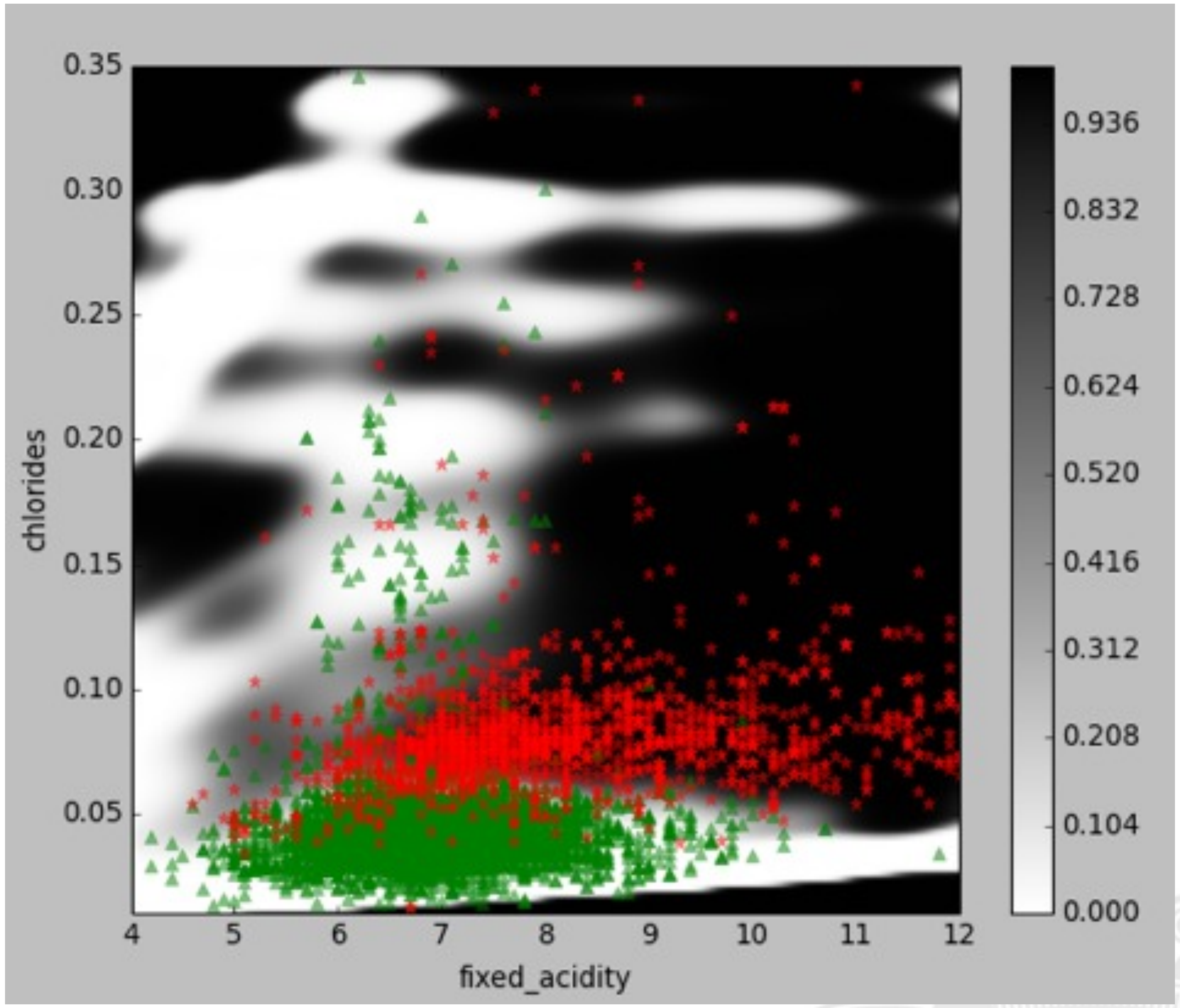
- **The second session: Model fitting**
  - Continue with the selected two variables
  - 1st model: A linear decision boundary
    - Visualize the result
  - 2nd model: A non-linear decision boundary
    - Create non-linearities using products and powers of the original variables
    - Visualize the result
  - See the script `hands-on.py`


0.35
0.30
0.25
0.20
0.15
0.10
0.05
chlorides
4
5
6
7
8
9
10
11
12
fixed_acidity
0.936
0.832
0.728
0.624
0.520
0.416
0.312
0.208
0.104
0.000

Reaktor

Model selection continues...

- How to tell whether a model is good or not?
- The figure on the right visualizes a decision boundary. Is it a good classification rule?
- How to select a model out of numerous possibilities?
- **The objective**: The model is supposed to *generalize* well, i.e. it should provide as accurate predictions as possible for *new* data (=data that was not used when fitting the model)



- Divide the data randomly to different parts:
  - **Training data set**: used to fit the model, i.e. optimize the parameters
  - **Validation/test data set**: used to evaluate prediction accuracy

Hands on continued

- **The third session: Model selection and evaluation**
  - Start by dividing the data set into two parts
  - Implement a process to test different models, e.g.
    - The best 2 variables out of available 11
    - With different nonlinear terms
    - The best K variables out of available 11
  - Find a model that generalizes well
    - Apply automated model selection
  - Hints in the end of the script `hands-on.py`

Summary

- A classification problem was introduced
- The session covered
  - Data insights
  - Model fitting
  - Model selection and evaluation
- In general, the presented process can be applied to any modeling task
- **Two possible solutions**
  - Using two variables with quadratic transformations: test accuracy 95%
  - Using all the available variables and fitting with regularization
    - The model uses a linear combination of 8 variables with *test accuracy 99.5%*
- Feedback: Github issue tracking or firstname.lastname@reaktor.fi