PTP SCIENCE PARK

Predicting mutation carriers from unbalanced data

Filippo Biscarini

H. Schwarzenbacher, H. Pausch, S. Biffani

Binary classification problems

Case/control

Disease diagnosis, response to treatments, susceptibility to diseases, survive or not ...

Sex

Male/female: e.g. sexed semen in cattle

Traceability

e.g. beef/non-beef meat

Mutations

Carriers/non-carriers (e.g. CVM in cattle)

Colour/breed

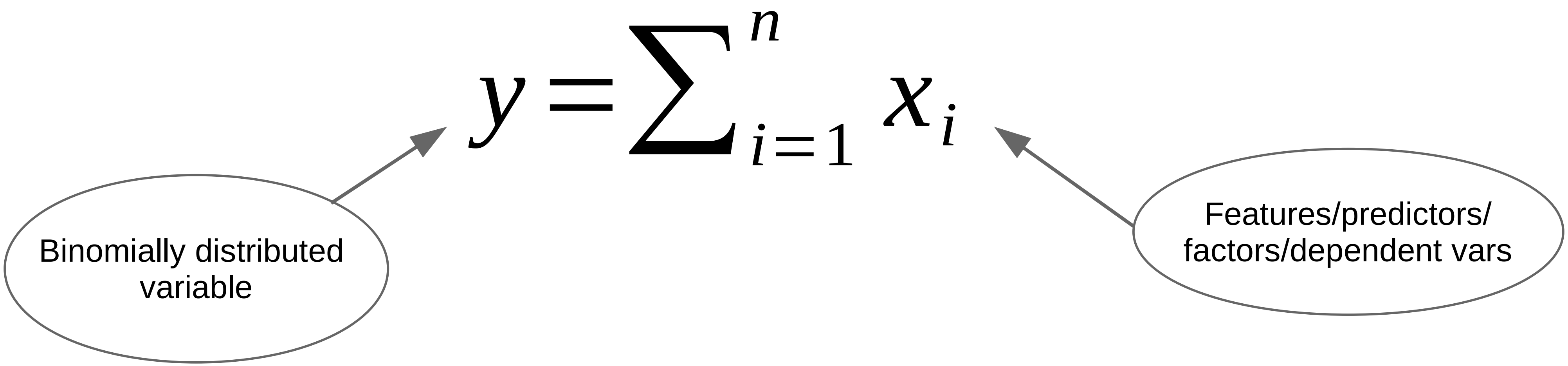
e.g. brown/white eggs; Pietrain/Landrace pigs

Gene alleles

e.g. casein variants in ruminants

Binary classification problems

- Collect binary observations
- Measure some quantities (on these obs) that are thought to be related to the binary outcome
- Model the outcome-features relationship



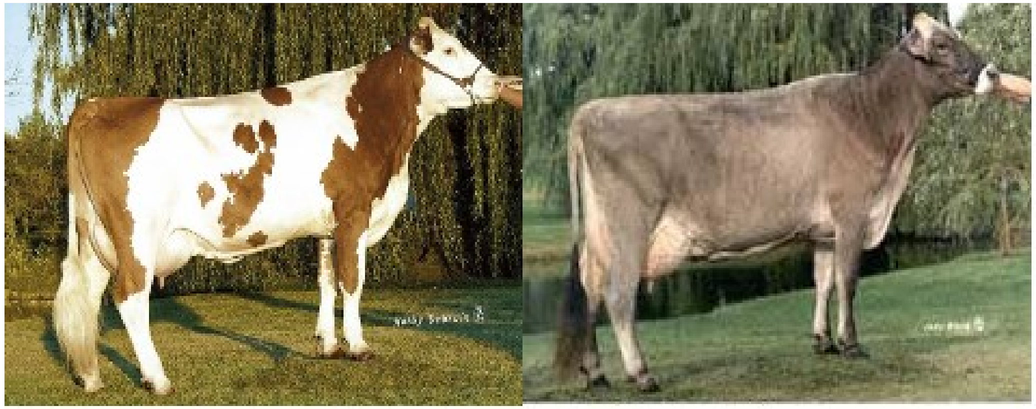
$$y=\sum_{i=1}^{n} x_i$$

Binomially distributed variable

Features/predictors/ factors/dependent vars

- Several methods available: logistic regression, (L)DA, SVM, KNN, classification trees ...

an illustration **from cattle genetics**



Mutation behind the **BH2** haplotype on **BTA19**
Two cattle breeds: **Brown Swiss**, **Fleckvieh**

3116 Fleckvieh: carriers/non-carriers: 126/2990
392 Brown Swiss: carriers/non-carriers: 250/142

SNP on BTA19: editing for call-rate (>95%)
Fleckvieh: 1317; Brown Swiss: 1370

Imputation (Beagle)

MAF: 0.224 in Fleckvieh, 0.187 in Brown

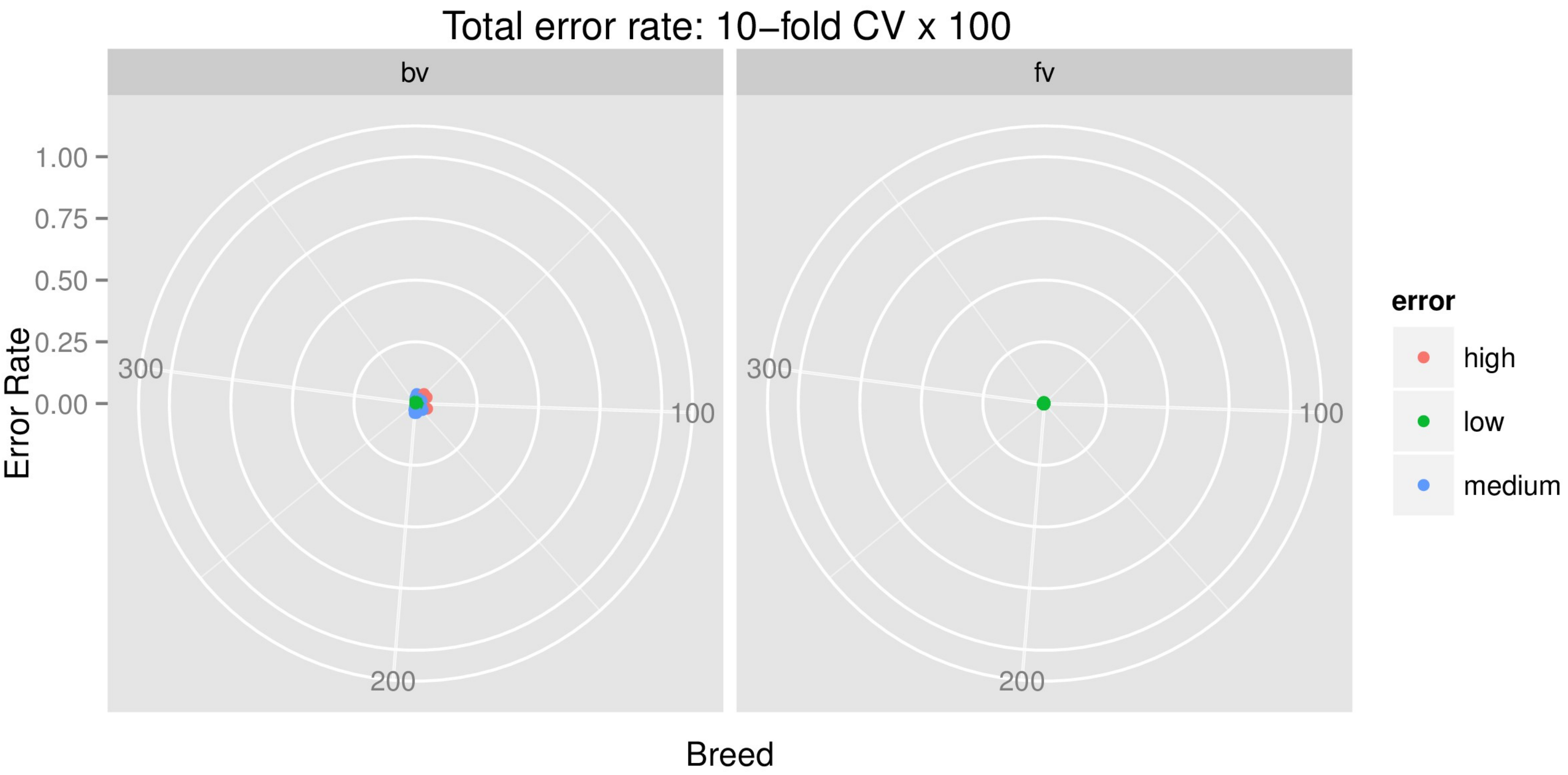
$$logit(p(x_i)) = \log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \mu + \sum_{j=1}^{m} z_{ij} SNP_j$$

Ridge logistic regression
(*p* > *n* in Brown Swiss)

80% data → training set: 5-gold CV to tune λ,
define the model
20% data → test set: estimate prediction accuracy

x 100 times

Total prediction accuracy



Fleckvieh: **99.78%** (± 0.2)

Brown Swiss: **98.91%** (± 1.1)

End of the story?

Extraordinarily effective classification!

Yes, **if** data were **balanced**

However:

Breed	**% carriers**	**% non-carriers**
Fleckvieh	4.04%	95.96%
Brown Swiss	63.78%	36.22%

Very **unbalanced data**, in opposite directions!

Classification with unbalanced data

Naive classifier: always predicts the majority class

Breed	E(Accuracy)
Fleckvieh	95.96%
Brown Swiss	63.78%

Beware: the accuracy in the minority class would be **0%**!

Not only total accuracy, but also accuracy in the two classes:

True positive rate: (identified carriers)/(all carriers)

True negative rate: (identified non-carriers)/(all non-carriers)

Classification with unbalanced data

Besides, what **type of error** is more **relevant**?

False positive or **false negative**?

- **False negatives**: critical in **recessive mutations**: more relevant to correctly **identify carriers** (to breed out), who could spread the defect
- **False positives**: **caseins**, better make sure that selected animals do **carry** the **positive variant**

Classification with unbalanced data

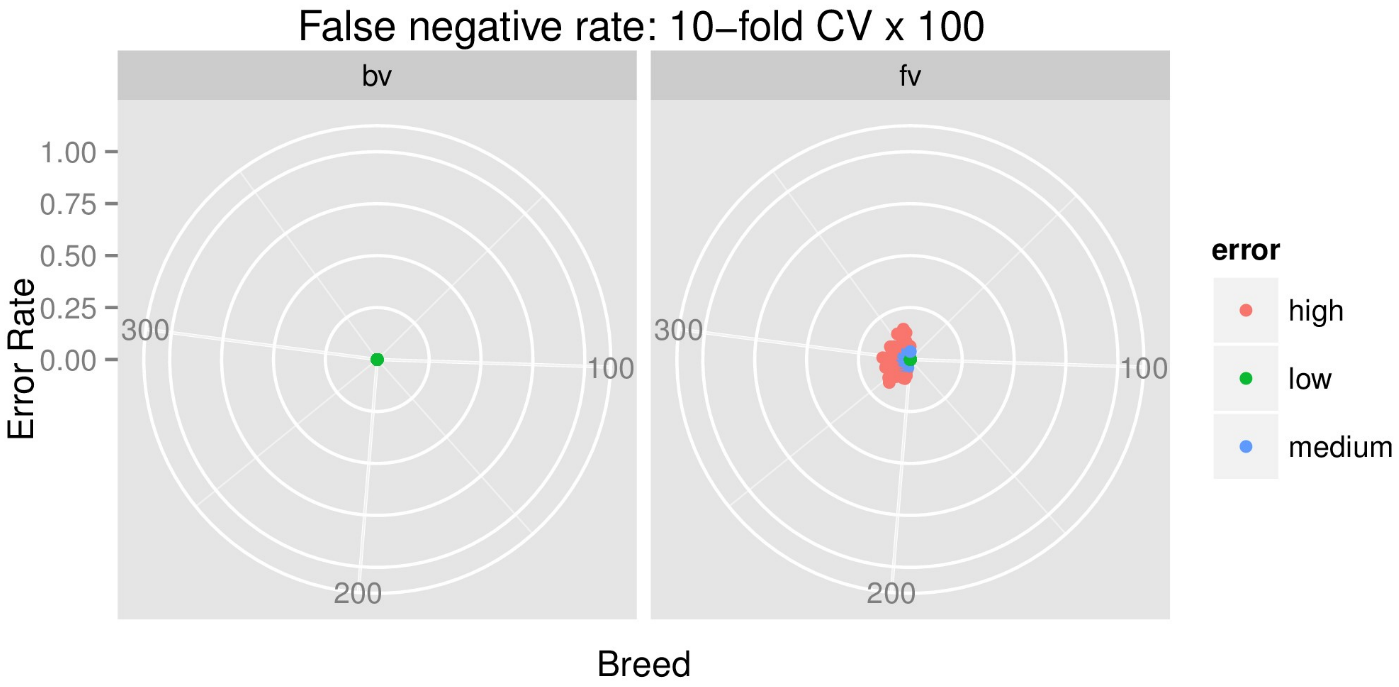
Besides, what **type of error** is more **relevant**?

False positive or **false negative**?

- **False negatives**: critical in **recessive mutations**: more relevant to correctly **identify carriers** (to breed out), who could spread the defect
- **False positives**: **caseins**, better make sure that selected animals do **carry** the **positive variant**

Animal geneticist's corollary to Murphy's law: the relevant case is always the minority class!

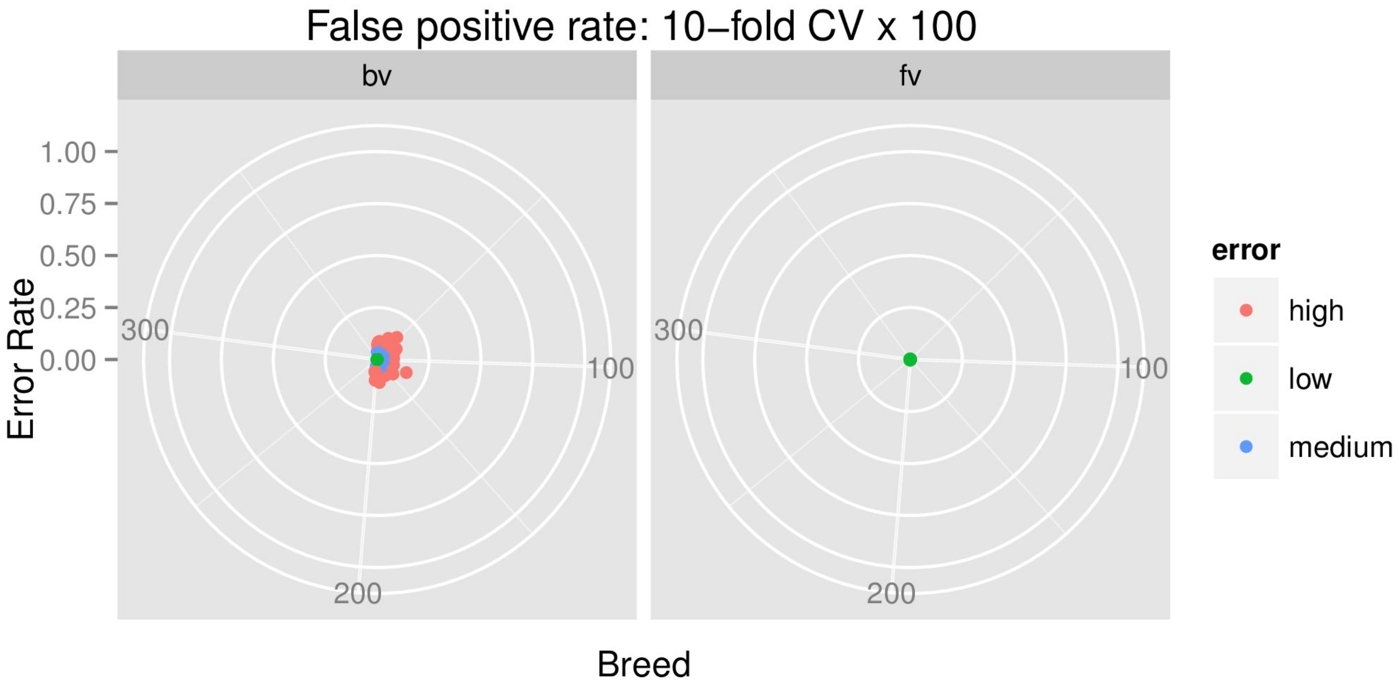
True positive rate



Fleckvieh: **95.51%** (± 3.67)

Brown Swiss: **100%** (± 0.0) [majority class!]

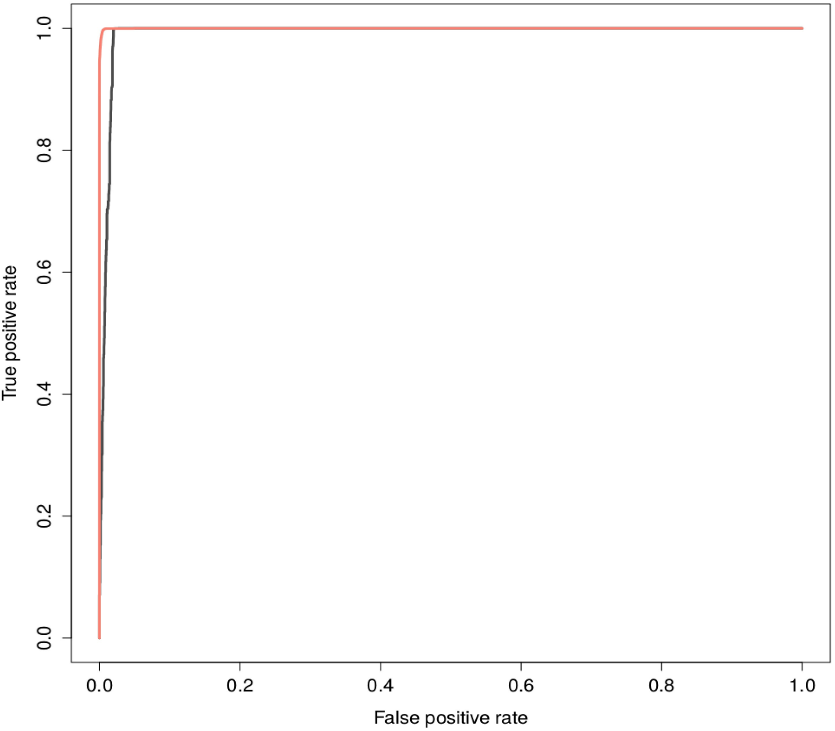
True negative rate



Fleckvieh: **99.95%** (± 0.08) [majority class!]

Brown Swiss: **96.96%** (± 3.07)

Dealing with unbalanced data



- Always look at the different types of errors!
- Try different classifiers → different TPR/TNR ratio
- Critically set the decision boundary
- ROC curves may help

- **Active learning**: design algorithm to optimize TPR/TNR in stead of overall accuracy [e.g. Ertekin et al., 2007]
- Sampling/re-sampling strategies: e.g. over- or under-sampling (informed or random)

- One-class learning [e.g. Tax, 2001]
- Ensemble methods like boosting may also help: combining several classifiers to improve classification performance

PTP
SCIENCE PARK

adding value **from research**

Thank you

www.ptp.it