Predictive ability metrics

Evangelina López de Maturana & Oscar González-Recio

Genome-wide prediction

Topics

Background

Overview of metrics

Metrics

R code

Why we need to evaluate the predictive ability of the models

Outcomes nature and metrics

Description of the metrics

Some examples

Why to evaluate the model performance?

A critical question in the genomic prediction is how accurately the genomic value predicts

It has to be informative to the potential end user of the model (e.g., breeder, clinician)

- AUC/c-index are among the most popular ones among clinicians
- Breeders: Mean squared error, Pearson's correlation, regression coefficient

Problems

Classification problems:

- analyzing medical data to determine if a patient is in a high risk group for a certain disease or not.

Regression problems

Time-to-event problems

Type of outcomes and metrics

- Continuous:
    - Distance/Bias
    - Mean Square Error
    - Pearson's, Spearman correlation
    - Coefficient of determination
    - Sensitivity, specificity, PPV, NPV, Type I error
- Binary:
    - Distance/Bias
    - AUC: area under the ROC curve
    - Coefficient of determination
- Survival outcome:
    - Distance/Bias
    - c-index

Bias/distance

Testing set: $y - \widehat{y}$

Continuous variables: y corresponds to the real outcome

Binary variables: $\widehat{y}$ corresponds to the predicted probability

Survival outcomes: $\widehat{y}$ is the predicted event probability at a given time

Better models have smaller distances between predicted and observed outcomes

Mean square error

The aim of any predictive model is to get as close as possible to the eventually realized value

The expected mean squared error of predictions can be equal to the squared predictions bias plus the variance of the prediction error

It measure stability and bias of the genomic proofs

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{CD_i} - \hat{y}_{CD_i}\right)^2$$

It may be expressed in units of standard deviations

Determination coefficient

The proportion of the total variance explained by predictors in the testing set:

$$R^2_{probit} = \frac{\mathrm{var}\left(\mathbf{X}_{test}\hat{\boldsymbol{\beta}}\right)}{\mathrm{var}\left(\mathbf{X}_{test}\hat{\boldsymbol{\beta}}\right) + \sigma_e^2}$$

Estimate of the h^2 (SNPs as predictors)

Sensitivity, specificity, PPV, NPV, Type I error

Sensitivity (True positive rate, Power): $\frac{TP}{TP + FN}$

Specificity (True negative rate): $\frac{TN}{TN + FP}$

Positive Predictive Value: $\frac{TP}{TP + FP}$

Negative Predictive Value: $\frac{TN}{TN + FN}$

Type I error: $\frac{FP}{FP + TN}$

Area under the receiver operating characteristic (ROC) curve

It provides information on the performance of classification models by balancing true positive (sensitivity) and false positive (1-specificity) discovering

ROC curve is a plot of the model sensitivity/TPR (y-axis) against the corresponding false-positive rate (1-specificity) (x-axis)

The curve is built from model performance at different thresholds

Area under the receiver operating characteristic (ROC) curve

It tells how much the model is capable of distinguishing between classes

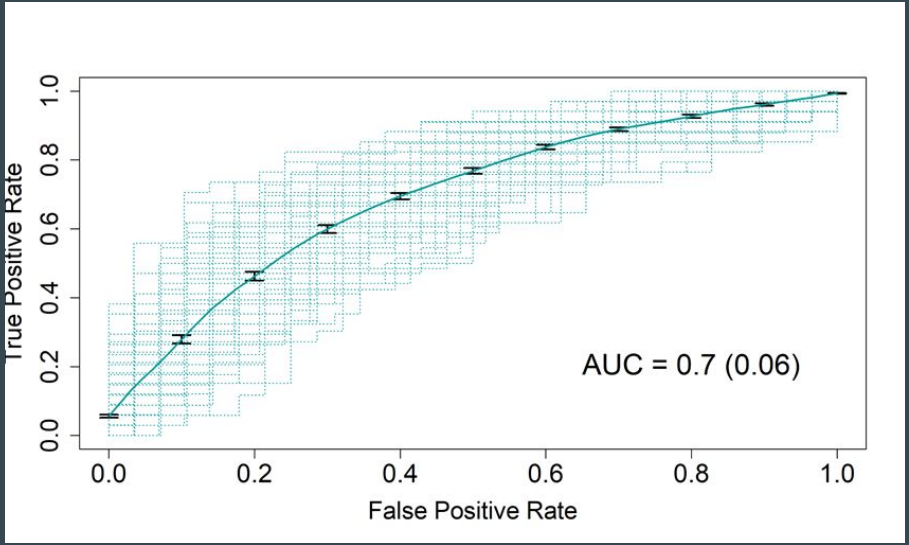
> the Higher the AUC, the better the model is at distinguishing between patients with and without the disease

AUC =1 → best classification performance; 0.5: the classification = random guess

Example: AUC=0.7 indicates that 70% of susceptible individuals present higher predicted liability than those non-susceptible

AUC

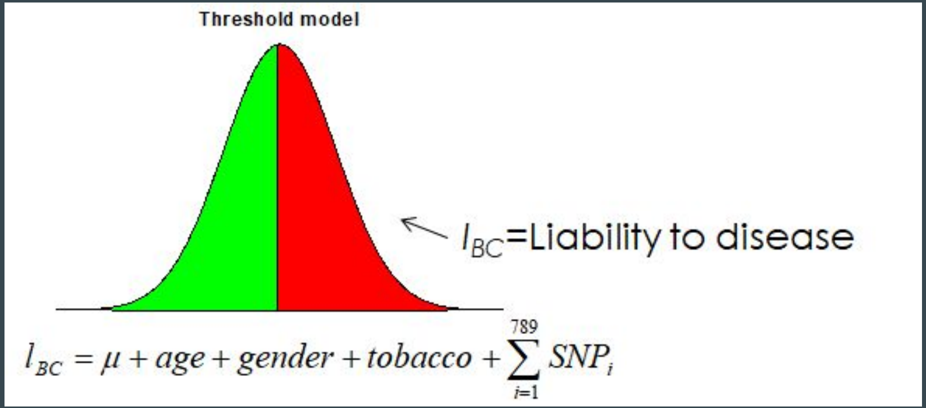
It provides an overall classification performance since AUC is averaged across all possible diagnostic thresholds



AUC

This criterion may be used for evaluating prediction models for categorical traits

Analysing binary traits may require to apply probit models, based on the assumption that a liability variable (instrumental variable) exists:



$$l_{BC} = \mu + age + gender + tobacco + \sum_{i=1}^{789} SNP_i$$

AUC

Validation set: It is not clear what threshold should be used to classify individuals, given their liability predictions

AUC may be useful

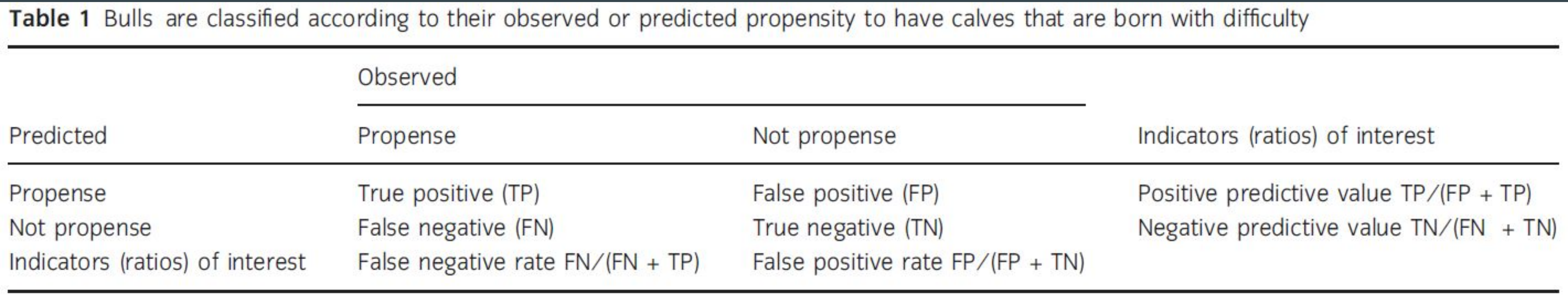
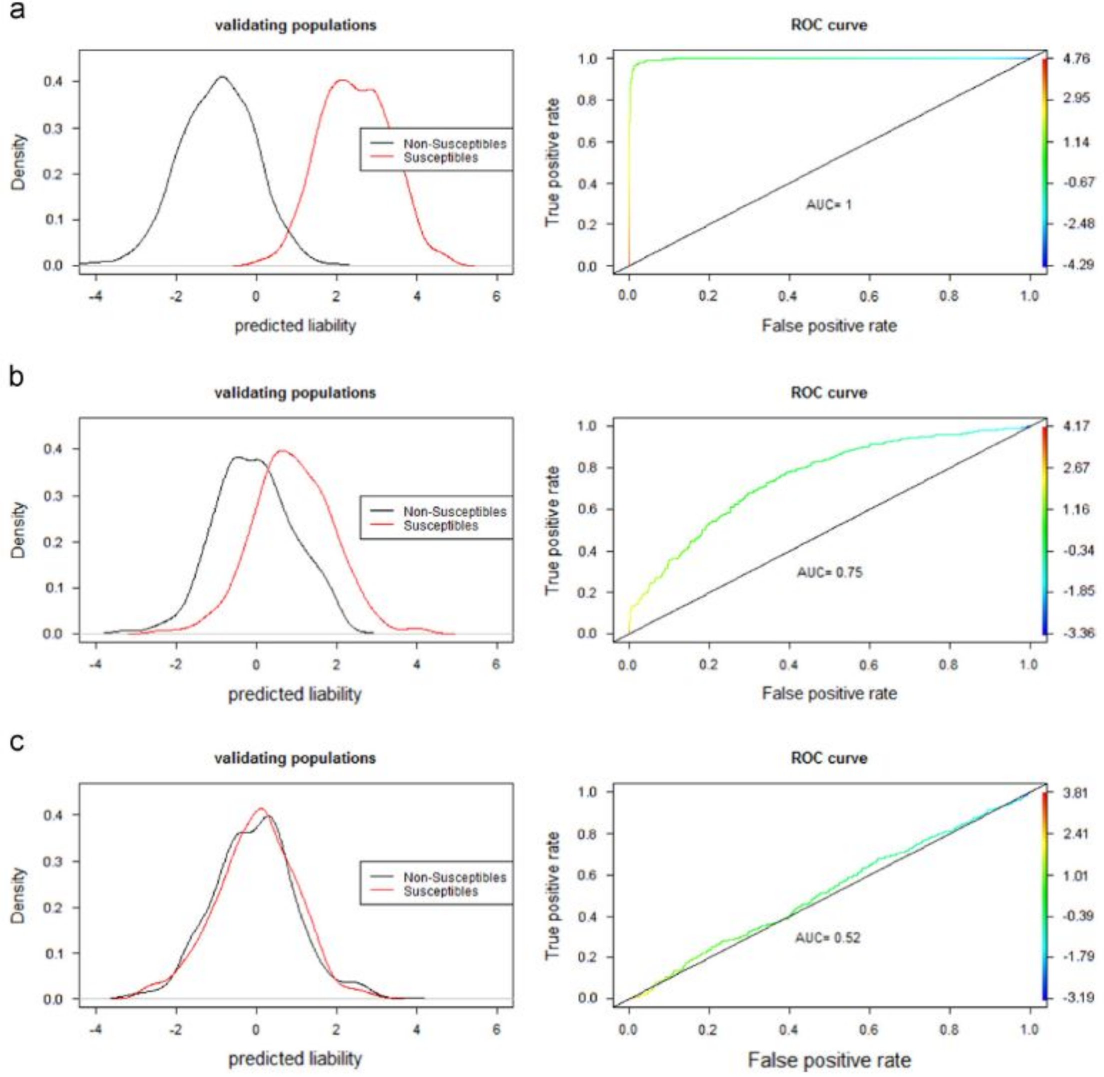


Table 1 Bulls are classified according to their observed or predicted propensity to have calves that are born with difficulty

	Observed		
Predicted	Propense	Not propense	Indicators (ratios) of interest
Propense	True positive (TP)	False positive (FP)	Positive predictive value TP/(FP + TP)
Not propense	False negative (FN)	True negative (TN)	Negative predictive value TN/(FN + TN)
Indicators (ratios) of interest	False negative rate FN/(FN + TP)	False positive rate FP/(FP + TN)	

López de Maturana et al., 2009



González-Recio et al., 2014

AUC - drawbacks

It does not differentiate between the accuracy with which the genomic profile predicts the true genetic risk of individuals and the accuracy with which true genetic risk predicts disease status (Wray et al., 2010)

It is determined by the heritability of the trait

It provides discrimination (separation of the classes) but not calibration (agreement between observed outcomes and predictions)

ROC AUC treats sensitivity and specificity as equally important overall when averaged across all thresholds

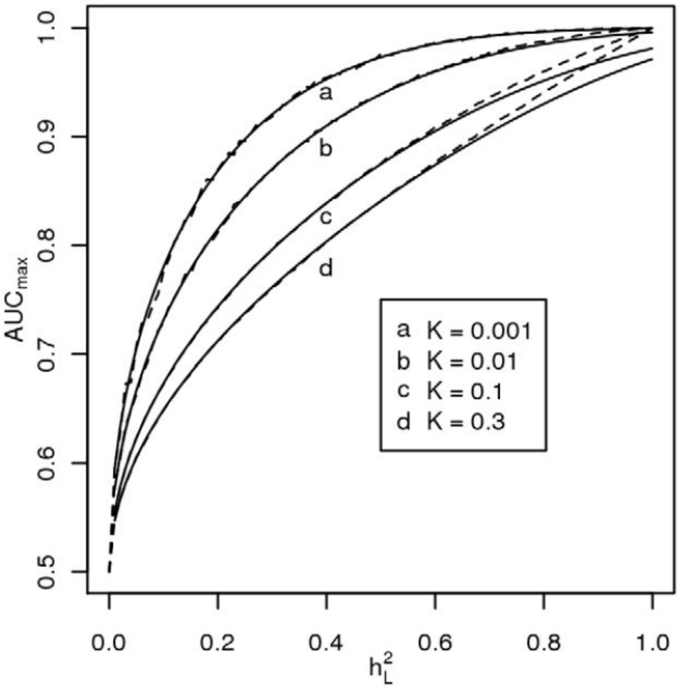


Figure 2. Relationship between maximum AUC (*AUC$_{max}$*) from a genomic profile and heritability on the liability scale h_L^2. For different disease prevalences (A–D) from simulation (dashed line) and from equation 3 (solid line).

doi:10.1371/journal.pgen.1000864.g002

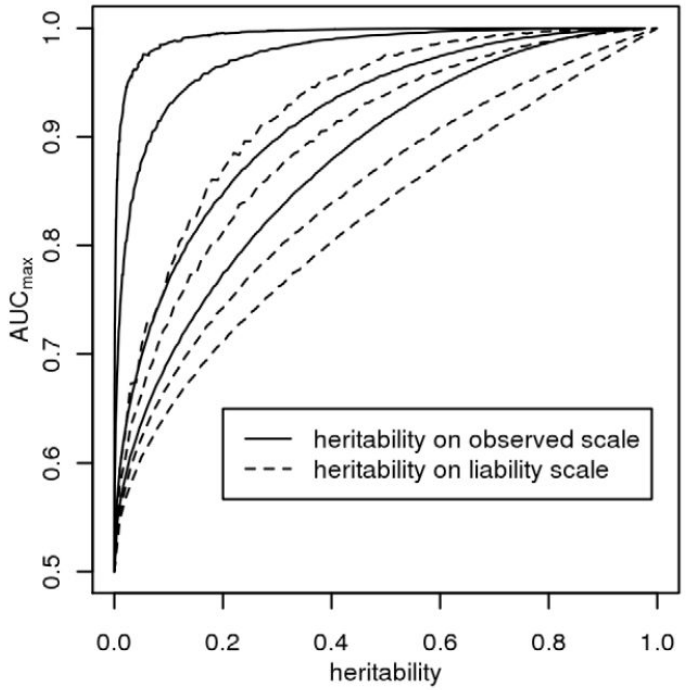


Figure 3. The relationship between maximum AUC (*AUC$_{max}$*) from a genomic profile and heritability on the liability scale h_L^2 (dashed line) or heritability on the observed scale H_{01}^2 (solid line), for disease prevalences in order from top left, *K* = 0.001, 0.01, 0.1, 0.3.

doi:10.1371/journal.pgen.1000864.g003

Concordance (c-) index

It is the generalization of the AUC that can take into account censored data

It measures the model’s ability to correctly provide a reliable ranking of the survival times based on the individual risk scores

Similarly to the AUC, *c*-index = 1 corresponds to the best model prediction, and *c*-index = 0.5 represents a random prediction

Drawback: it does not provide a value specific to the time horizon of prediction

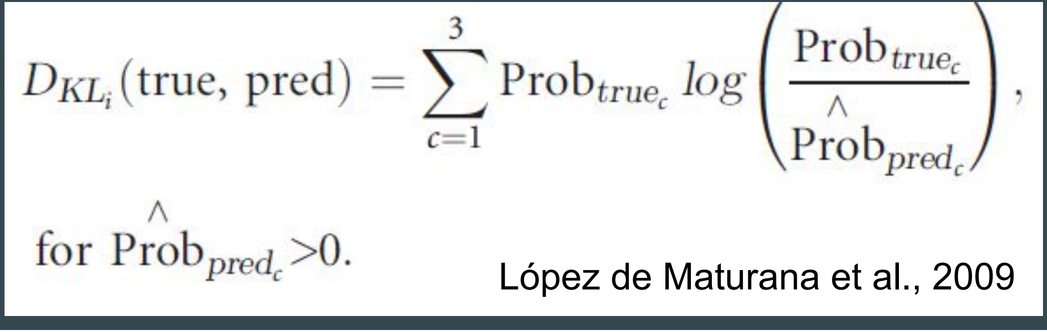
Kullback-Leibler distance

Measure of the difference between two probability distributions, true (in this case, the observed data distribution) and the alternative (predicted distribution in the testing set)

Models with the smallest K-L distance would be favoured

Outcome with 2 or more categories

The K-L for each data point *i*:



$$D_{KL_i}(\text{true, pred}) = \sum_{c=1}^{3} \text{Prob}_{true_c}\, log\left(\frac{\text{Prob}_{true_c}}{\overset{\wedge}{\text{Prob}}_{pred_c}}\right),$$

for $\overset{\wedge}{\text{Prob}}_{pred_c} > 0.$

López de Maturana et al., 2009

Pearson’s correlation

Covariance between observed and predicted outcomes

$$\rho_{y_{CD}\hat{y}_{CD}} = \frac{cov(y_{CD}, \hat{y}_{CD})}{\sigma_{y_{CD}}\sigma_{\hat{y}_{CD}}}$$

Standard deviation for the observed outcomes

Standard deviation for the predicted outcomes

Spearman’s correlation

When apply GS, we are interested in ranking the individuals in order to select the best ones as the parents of future generations (animal and plant breeding)

We may want to compare if the rankings provided by two GS procedures differ

It is a nonparametric measure of rank correlation

Some tips

None of the metrics provide a full representation of predictive ability

Better to use several criteria

Topics

Background

Overview of metrics

Metrics

R code

Why we need to evaluate the predictive ability of the models

Outcomes nature and metrics

Description of the metrics

Some examples

Hands-on

4_Metrics.R

5_AUC.R