Testing the Significance of Attribute Interactions

Aleks Jakulin & Ivan Bratko

1st of July, 2004

1. OBJECTIVES

- Quantifying *interactions* among k attributes?
- Using *information-theoretic* measures to quantify relationships and relations between attributes?
- Is an attribute *significantly* associated with the label? Do attributes interact significantly?
- Comparison between the χ^2 distribution, *bootstrap* and the *cross-validation*.

2. Quantifying Interaction

Attributes $\mathcal{V} = \{X_1, \ldots, X_k\}$ *interact* if the *whole* probability mass function $P(\mathcal{V})$ cannot be reconstructed from *parts*: $\{P(X_2, X_3, \ldots, X_k), P(X_1, X_3, \ldots, X_k), \ldots, P(X_1, X_2, \ldots, X_{k-1})\}$

Heuristic: A heuristic estimate shown to be useful for quantifying interaction in practice is McGill's interaction information:

$$I(X_1; X_2; \ldots; X_k) = I(\mathcal{V}) = -\sum_{\mathcal{T} \subseteq \mathcal{V}} (-1)^{|\mathcal{V}|-|\mathcal{T}|} H(\mathcal{T}).$$

Meaning: Interaction information is the loss (relative entropy) caused by approximating the whole from the parts using the generalized Kirkwood superposition approximation.

$k = 2$: Mutual Information

Whole: $P(A, B)$ **Parts:** $\{P(A), P(B)\}$

The approximation from parts is $\hat{P}(A, B) = P(A)P(B)$. To compute the approximation error, use the *Kullback-Leibler divergence* or relative entropy:

$$D(P||\hat{P}) = \sum_{a,b} P(a, b) \log \frac{P(a, b)}{\hat{P}(a, b)} = I(A; B) = H(A)+H(B)-H(A, B)$$

This corresponds to mutual information (information gain). Also, this corresponds to the loss made in conditional prediction, making mutual information relevant also for *supervised learning*:

$$I(A; B) = D(P(A|B)||P(A)) = D(P(B|A)||P(B)).$$

If the loss is high the attribute is relevant for predicting the label.

$k = 3$: Interaction Information

Whole: $P(A, B, C)$ **Parts:** $\{P(A, B), P(A, C), P(B, C)\}$

The chain rule does not help, because the dependencies are cyclic.

The closed-form solution is the *Kirkwood superposition approximation*:

$$\hat{P}(A, B, C) = \frac{P(A, B)P(B, C)P(A, C)}{P(A)P(B)P(C)} = P(A|B)P(B|C)P(C|A)$$

It is a special case of Kikuchi and for $k = 3$ of mean-field approximations. It is also relevant as an approximation of the loss made by the naïve Bayes in supervised prediction:

$$D(P||\hat{P}) = I(A; B; C) = D\left(P(C|A, B)\middle\|\frac{P(C)P(A|C)P(B|C)}{P(A)P(B)}\right) =$$

$$= H(A, B) + H(A, C) + H(B, C) - H(A) - H(B) - H(C) - H(A, B, C)$$

Interaction information:

- is symmetric: $I(A;B;C) = I(B;A;C) = I(C;B;A) = \ldots$.
- corresponds to the influence of one attribute on the mutual information between the other two: $I(A;B|C) - I(A;B)$.
- corresponds to the difference between the mutual information between the label and both attributes *together* versus *apart*: $I(AB;C) - (I(A;C) + I(B;C))$.

Interaction information can be:

- **POSITIVE:** There is a pattern among the k attributes unlocked only in the presence of them all simultaneously $I(AB;C) > I(A;C) + I(B;C)$ *synergy*.
- **ZERO:** There is no pattern of order k: $I(A;B|C) = I(A;B)$.
- **NEGATIVE:** There is duplication among the parts $I(AB;C) < I(A;C) + I(B;C)$ *redundancy*.

3. Applications of Interactions

The relationship between a set of attributes is worth considering if and only if the attributes interact.

(Otherwise, the relationship is deducible from simpler relationships in the parts.)

Examples of relationships:

- Similarity between attributes $I(A;B)$ and the relevance for classification/regression $I(A;Y)$.
- Deciding which attributes should be combined when predicting the label $I(A;B;Y)$.
- Examining a scatter plot, rule induction, constructive induction.

Methodological Notes

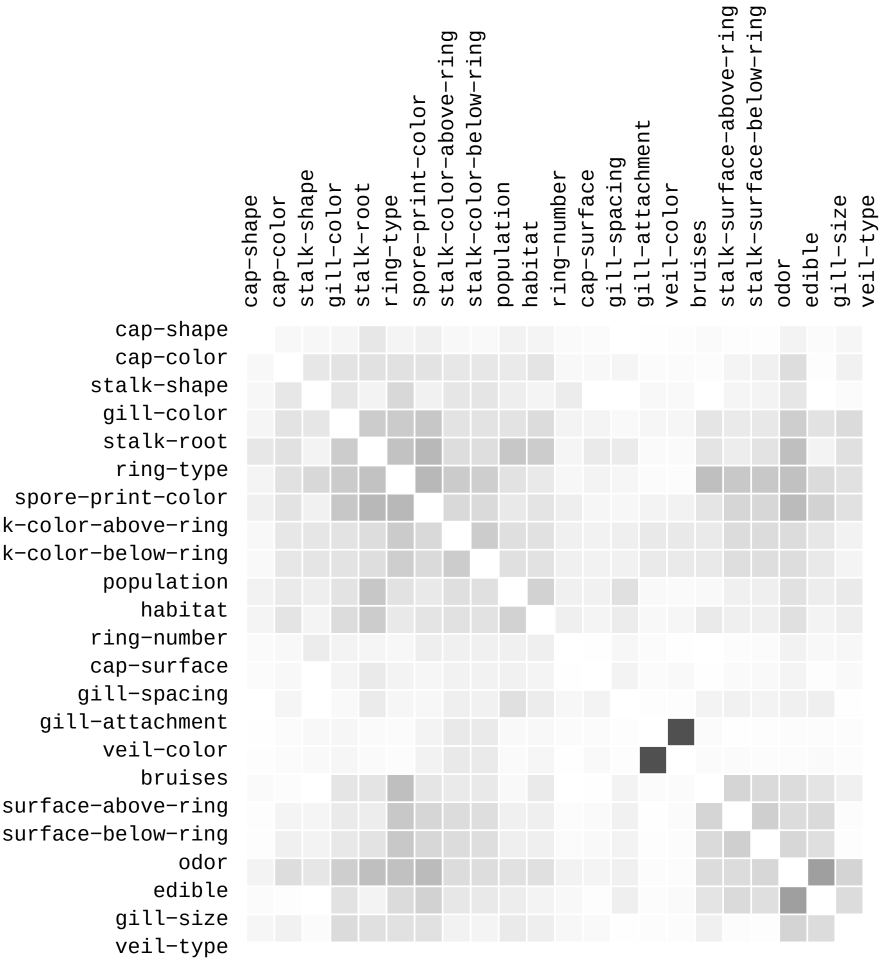
- Relative frequencies are used to estimate the joint probability mass function $P(A, B, C)$. Bayesian priors only add bias, but no benefit for this purpose.

- We use color to convey the type of the interaction: POSITIVE - synergy, ZERO - no interaction, NEGATIVE - redundancy.

- In supervised learning tasks, we can express interaction information as a percentage of the label entropy $H(Y)$. In unsupervised learning tasks, the interaction information can be expressed as the percentage of the joint entropy $H(A, B, C)$. We refer to these as *normed* interaction information.

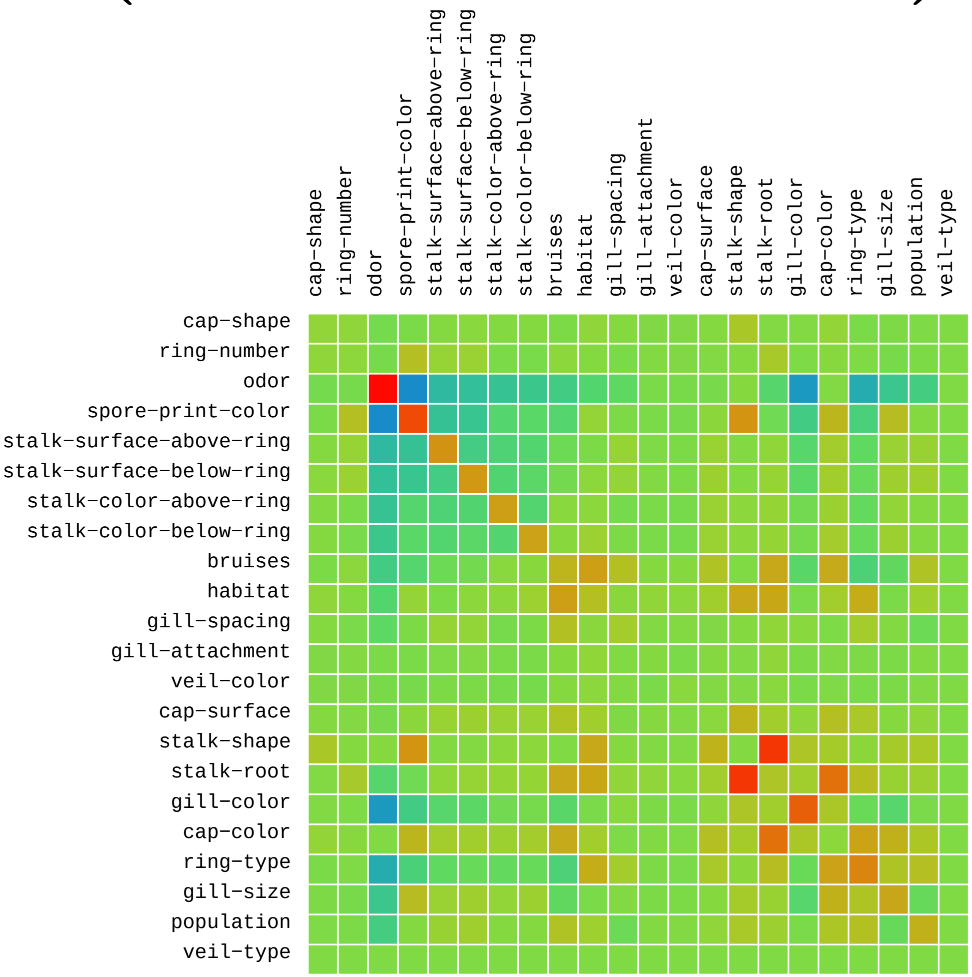
- The corresponding proximity measure is the absolute value of the normed interaction information, and ranges from 0 to 1. The similarity measure derived from this $(1 - d)$ obeys the triangle inequality.

Interaction Matrices (Mushroom Data)



2-way interactions

The darker the rectangle the higher the mutual information $I(A;B)$ between two attributes.

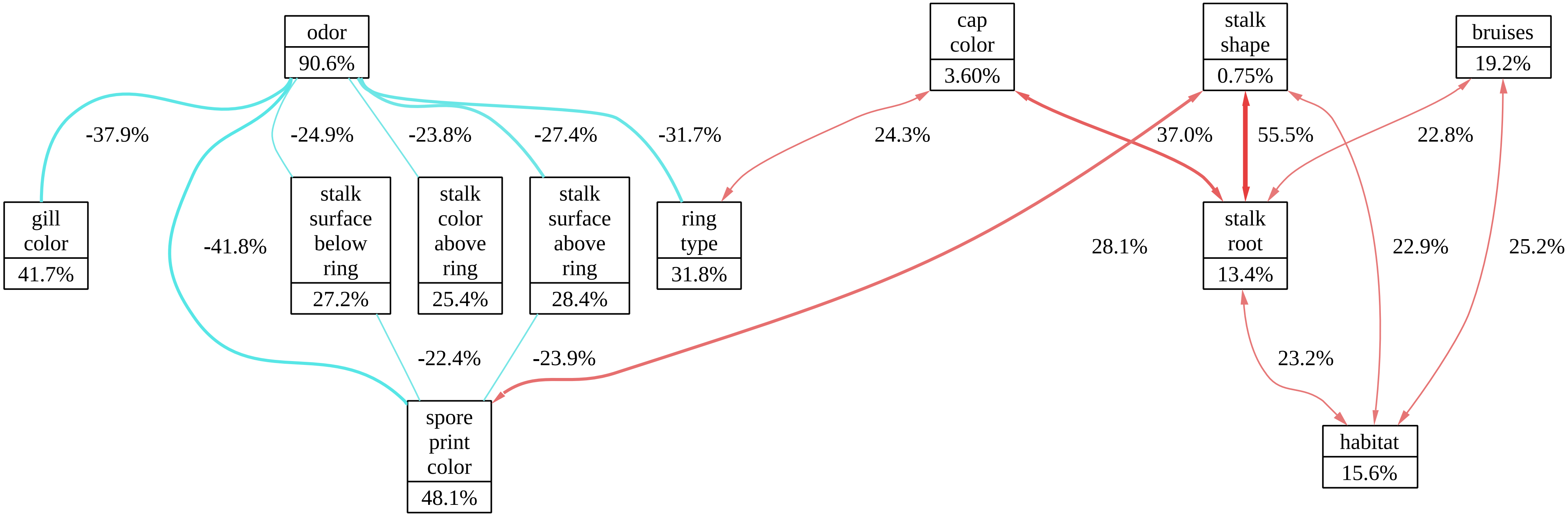


3-way interactions

The bluer/redder the rectangle the higher the $I(A;B;Y)$, where Y is always the label, ‘edibility’.

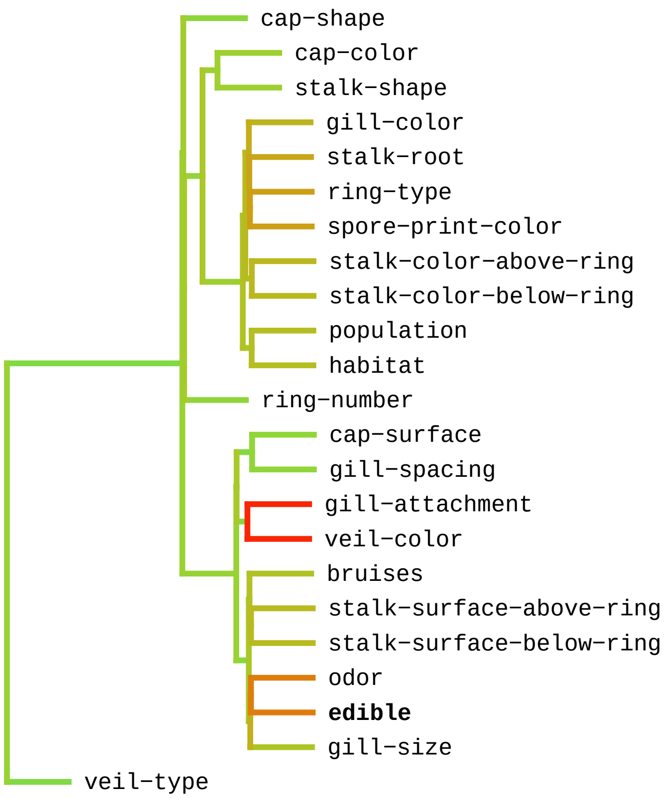
Interaction Graph (Mushroom Data)

The most distinct 3-way interactions can be shown as a graph, which can be seen as a summary representation of the interaction matrix:

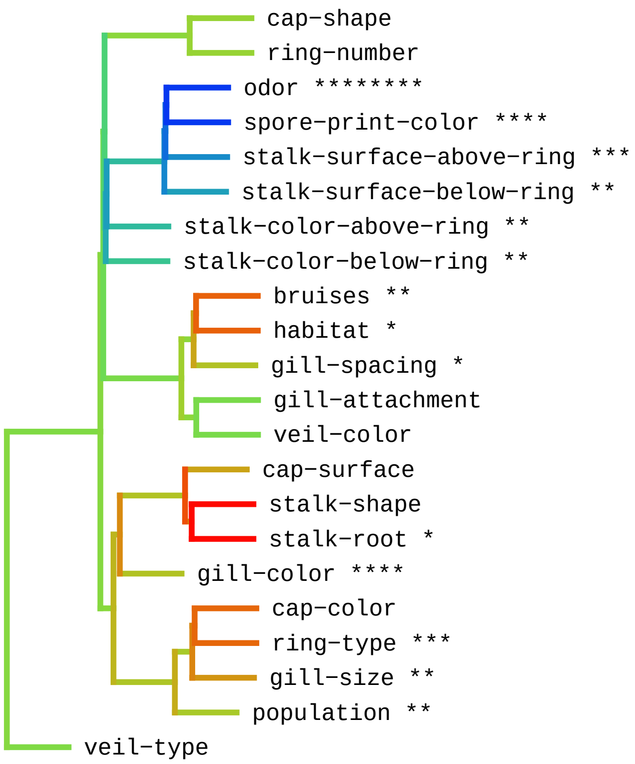


The percentages are expressed as a percentage of the label entropy $H(Y)$. Conditional mutual information can be obtained from this graph easily: given 'odor' the 'gill color' attribute only provides $41.7 - 37.9 = 3.8\%$ of information about the label; this attribute is conditionally irrelevant.

Interaction Dendrograms (Mushroom)



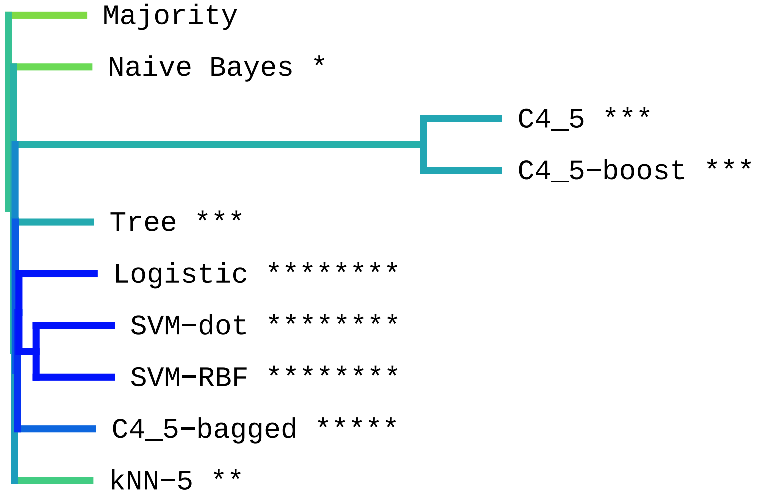
2-way interactions

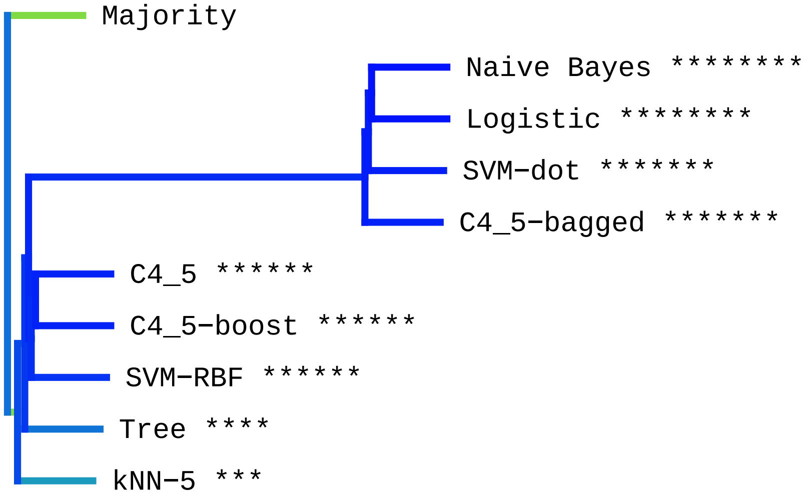


3-way interactions

Another summarized form of the matrix is the dendrogram. Asterisks ‘*’ indicate the mutual information between the attribute and the label.

Learning Algorithm Taxonomies

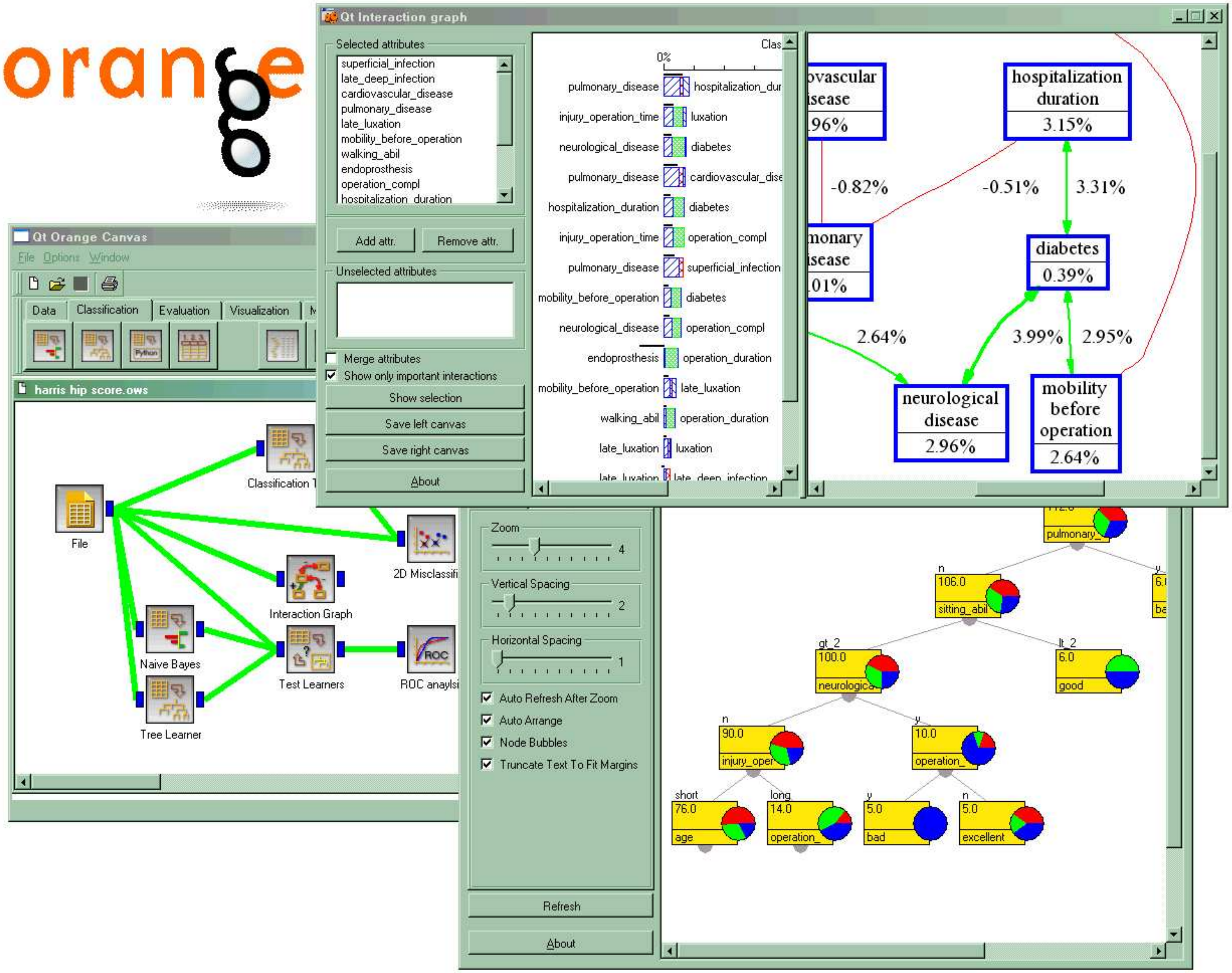




Tic-Tac-Toe **Pima-Diabetes**

Taking a classifier as an attribute, we can use interaction dendrograms for creating a taxonomy of classifiers based on comparing the classifiers through their interactions. We can see which classifiers are best (asterisks), which of them are correlated and which are complementary.

Orange - Interactive Interaction Analysis



☐ Merge attributes

☑ Show only important interactions

☑ Auto Refresh After Zoom

☑ Auto Arrange

☑ Node Bubbles

☑ Truncate Text To Fit Margins

Download from `http://www.ailab.si/orange` (GPL, Python).

4. Part-to-Whole Models

Interaction information is based on Kirkwood superposition approximation that does not always yield a proper probabilistic model when $k > 2$: it may not not sum to 1. The resulting *negative* losses are meaningless, but can be remedied by normalization.

Alternatively, a different part-to-whole approximation may be used. A good choice are maximum entropy models constrained by the parts. This formulation is not in closed form (except in certain cases) and requires an iterative optimization procedure, such as *generalized iterative scaling*. This underlies the definition of interaction as given by I. J. Good in 1963.

Two Kinds of Normalization

- *Joint* normalization:

$$\hat{P}'(a, b, c) = \frac{\hat{P}(a, b, c)}{\sum_{a', b', c'} \hat{P}(a', b', c')}$$

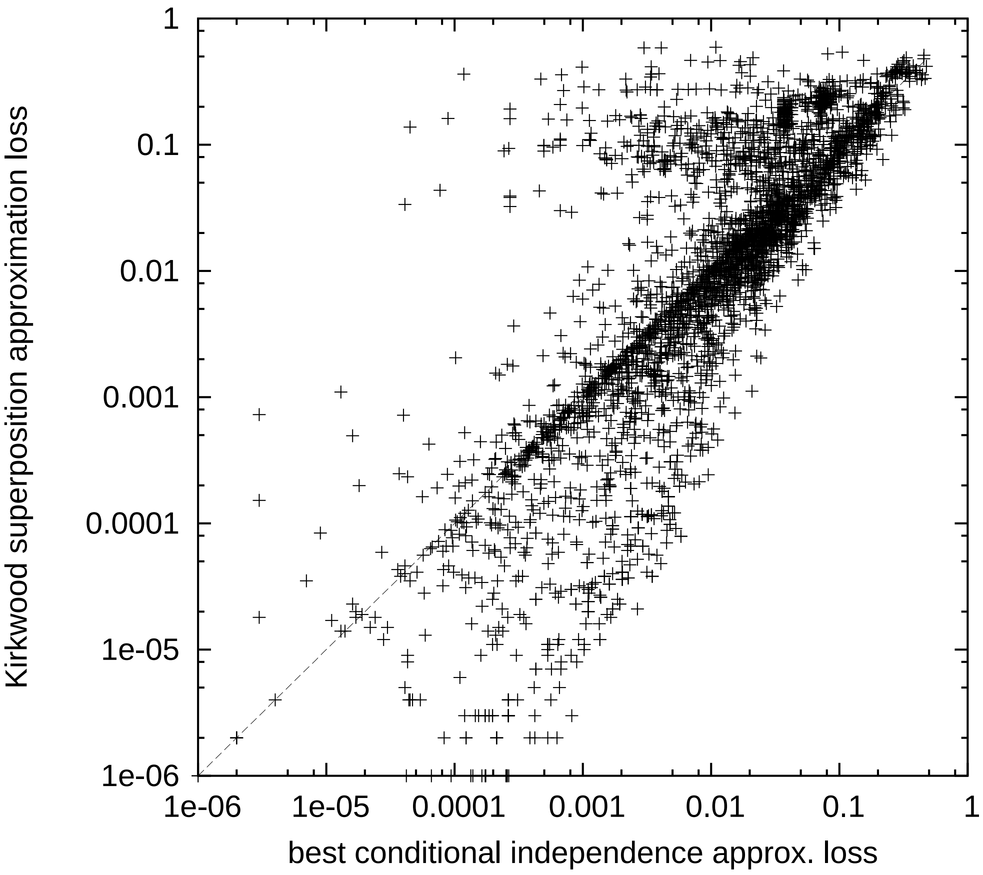
- *Conditional* normalization:

$$\hat{P}'(a|b, c) = \frac{\hat{P}(a, b, c)}{\sum_{a'} \hat{P}(a', b, c)}$$

These two methods are generally different. In the remainder of the experiments, we will be performing unsupervised normalization. The loss will be always computed as Kullback-Leibler divergence $D(P||\hat{P}')$. *Before computing the KL-divergence, $\hat{P}$ must be normalized!*

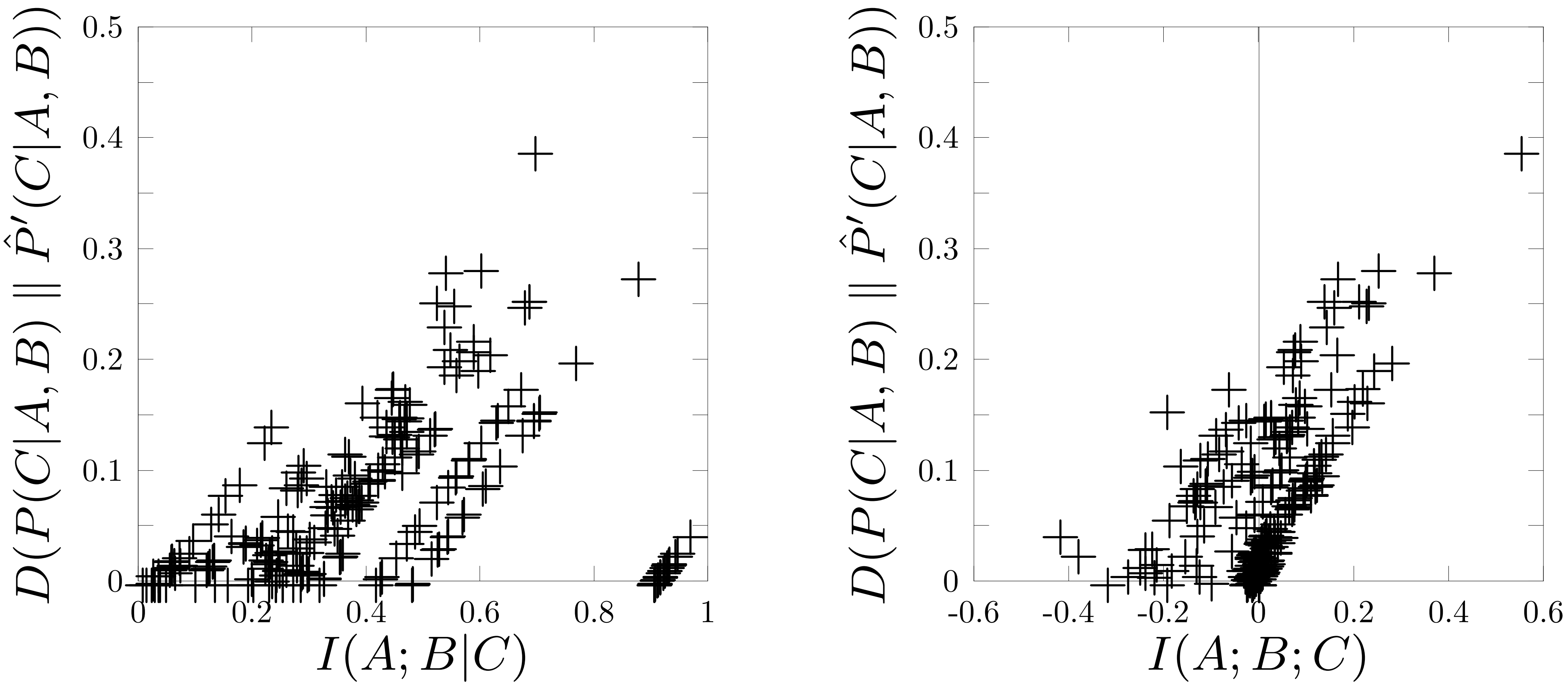
Can Adding Parts Increase the Loss?

It may seem that the more parts there are, the lower the loss. Unfortunately, Kirkwood superposition approximation (KSA) does not have this property. We see this by comparing KSA with models that assume conditional independence.



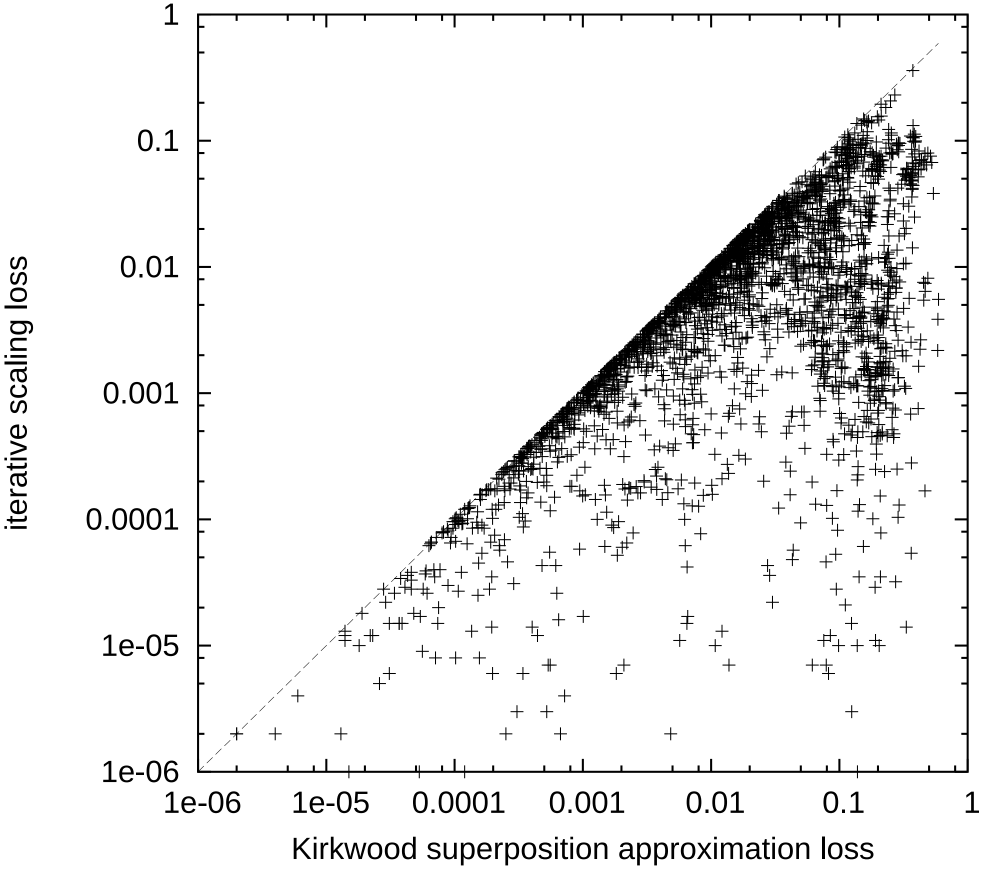
Predicting the Loss of Naïve Bayes

Conditional mutual information $I(A;B|C)$ is often used as the predictor of the loss caused by the conditional independence assumption in the naïve Bayesian approximation $\hat{P}'(C|A,B)$. Interaction information $I(A;B;C)$ works better (Mushroom data):



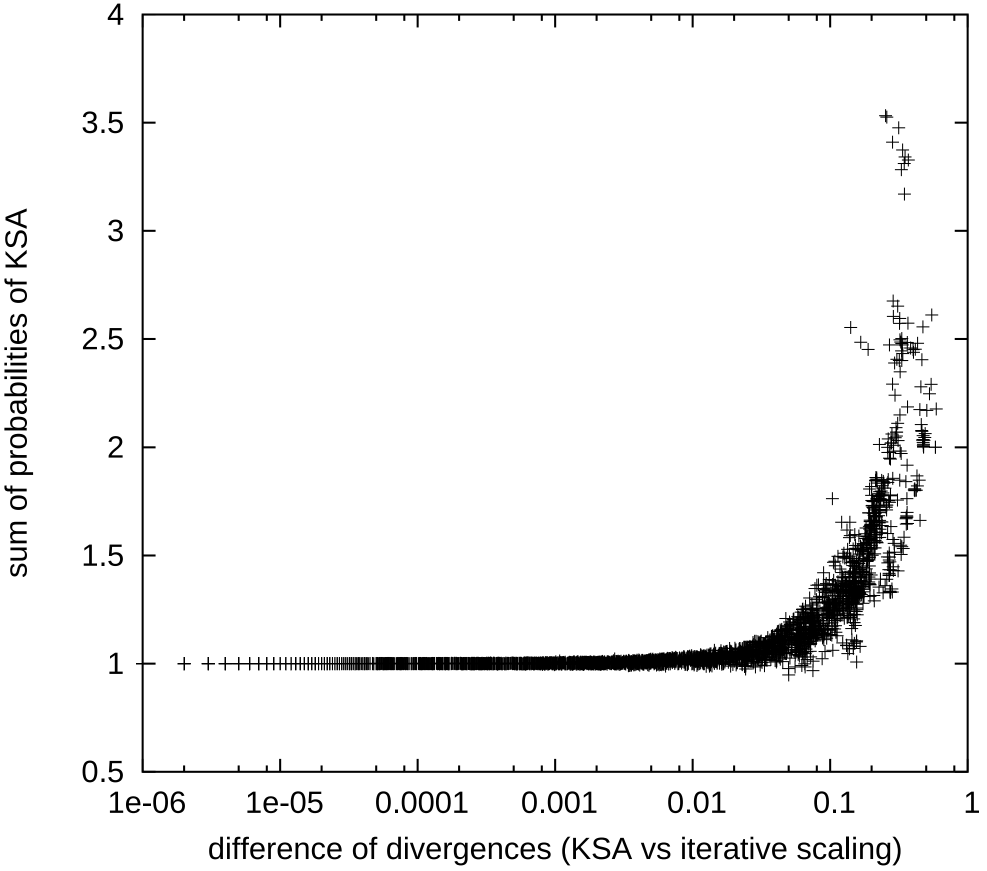
Iterative Scaling vs. KSA

Iterative scaling always achieves better results than the Kirkwood superposition approximation. As such, it is a better part-to-whole approximation (but it is not in closed form).



Normalization is Imperfect!

We can predict the loss caused by the Kirkwood superposition approximation in comparison to iterative scaling. When the normalization factor is very different from 1, the iterative scaling is practically guaranteed to be much better.



5. Significance Testing

The essential element of significance testing is the realization that even the *correct model usually incurs loss* on a finite sample from the model itself.

The probability distribution of this *self-loss* provides a realistic scale in which the *approximation loss* can be described as a probability that the self-loss is greater or equal to the approximation loss. This probability is the *P-value*.

Important: P-values depend on the part-to-whole approximation used. We employ Kirkwood superposition approximation, but all the testing methods are independent of the approximation method.

Test-Bootstrap Protocol

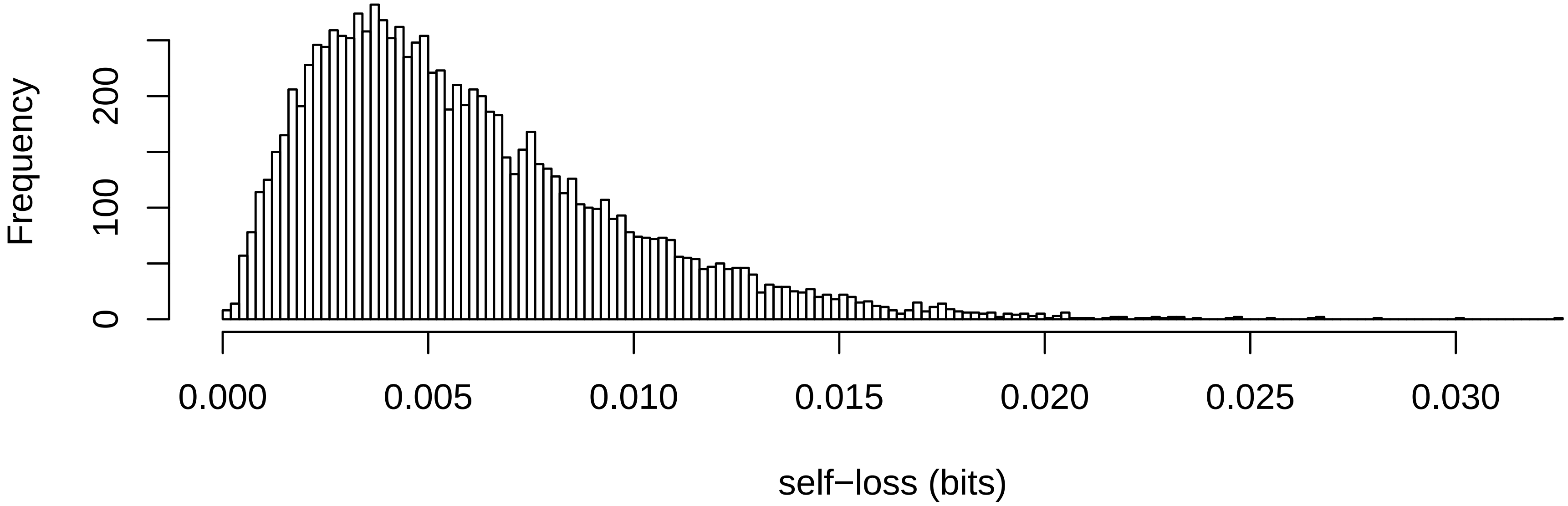
We obtain the self-loss distribution by *perturbing the test data*. We are not interested in the generalization power (to unseen data), and we are not interested in the distribution of the approximate model.

1. Using all the data $\mathcal{D}$, compute the approximate model $\hat{P}(X|\mathcal{D})$.
2. Using $\mathcal{D}$, compute the max. likelihood reference model $P(X|\mathcal{D})$.
3. Create a bootstrap resample $\mathcal{D}'$ of the same size from $\mathcal{D}$.
4. Compute the max. likelihood reference resample model $P'(X|\mathcal{D}')$.
5. The P-value equals $\Pr\{D(P'(X)||P(X)) \geq D(P(X)||\hat{P}(X))\}$.

Beware: The procedure differs from the 'usual' bootstrap procedure which perturbs the training set.

Self-Loss

The distribution of self-loss ($D(P'||P)$) over 10000 bootstrap resamples in ‘voting’ data for the *immigration* attribute:



Mutual information ($D(P||\hat{P}) = 0.002467$ bits) indicates that the *P-value is 0.6986*. In 70% of the experiments, the self-loss was larger than the approximation loss. This interaction is insignificant, meaning that the variation in the performance exceeds the approximation error.

χ^2, Goodness-of-Fit, and KL-Divergence

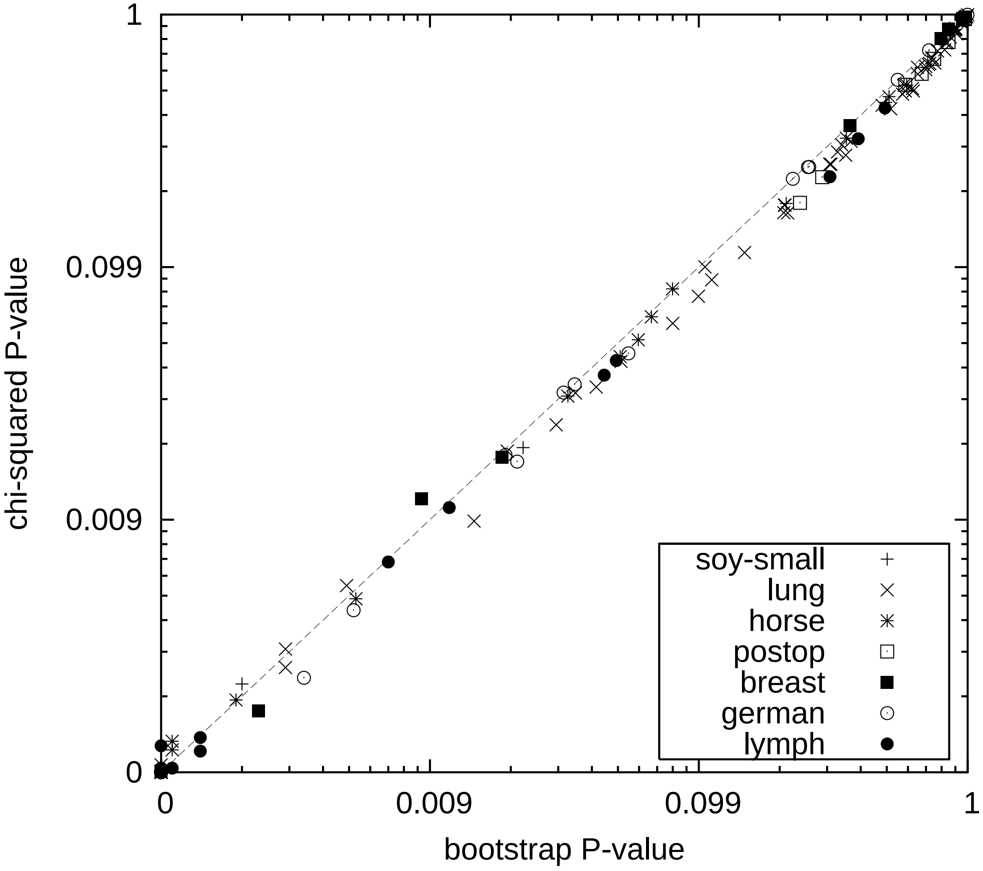
If the underlying reference model $P(\mathcal{V})$ is based on relative frequencies estimated from n instances, the KL-divergence between P and an independent joint PDF $\hat{P}$ multiplied by $2n/\log_2 e$ is equal to the Wilks' likelihood ratio statistic G^2. In the context of a goodness-of-fit test for large n, G^2 has a χ^2_{df} distribution with *df* degrees of freedom:

$$\frac{2n}{\log_2 e} D(P \| \hat{P}) \underset{n\to\infty}{\sim} \chi^2_{|\Re_{\mathcal{V}}|-1} \tag{1}$$

Here, $df = |\Re_{\mathcal{V}}| - 1$ is based on the cardinality of the set of possible combinations of attribute values $|\Re_{\mathcal{V}}|$. (Certain value combinations may not appear in the data, and the *df* must be correspondingly lower. Zero counts don't count as degrees of freedom.)

χ^2 vs. Test-Bootstrap

Self-loss as assessed by test-bootstrap protocol and multiplied by $2n/\log_2 e$ has a χ^2 distribution asymptotically as $n \to \infty$. But how about small data sets with n in the range 30-1000? *YES, almost identical!*



The Principle of Occam's P-Razor

Pick the simplest model among those that are not significantly worse than the best one.

P-values Questions & Answers:

- In larger data sets, the variation in self-loss becomes very small: everything is significant. **Solution:** use smaller resamples, as the improvement should be significant even there.
- Multiple-testing problem. **Solution:** study the correlation between P-values, don't assume it.
- I care about expected performance, not about simplicity! **Solution:** These two criteria are different, but risk does matter.
- I care about truth, not about loss. **Solution:** Your utility function is the model's *a posteriori* likelihood, given the data.

6. Cross Validation

P-values ignore both the variation in approximation loss and the generalization performance of a model. CV-values are a solution.

CV-values evaluate the *probability* that one model is better than another on separate test data. We are *not* interested in the expected performance difference, but in the *distribution* of performance difference. Namely, P-values may underestimate the quality of approximate models, which often have lower performance variation (model variance) than more complex models.

Cross-Validation Protocol

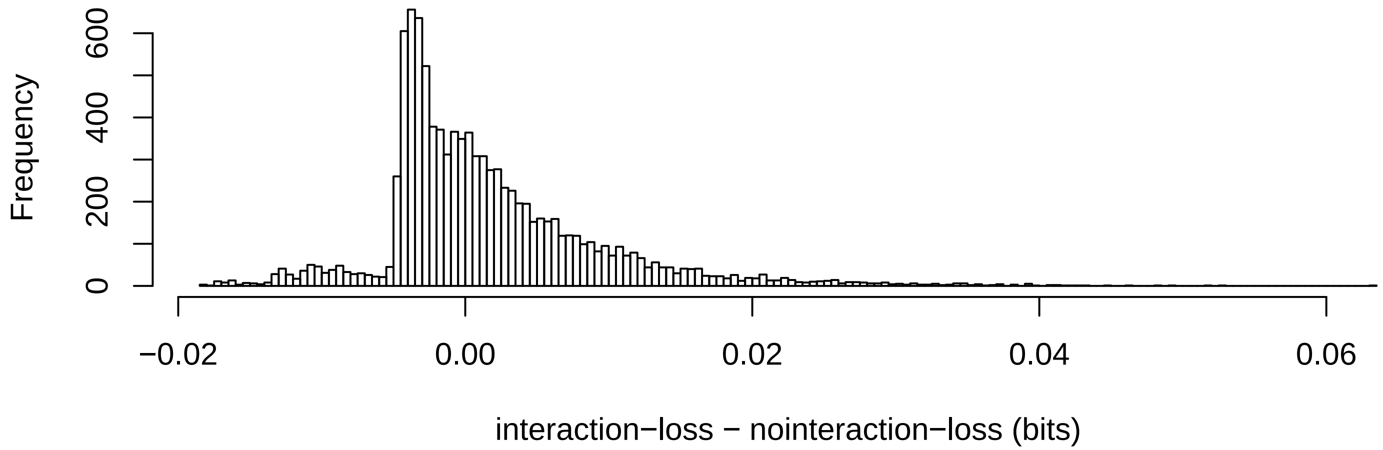
Since probability does not exist without a model, the same trusted reference model is trained both on training and the test data. The most meaningful number of folds is therefore ~ 2. The reference model should make a honest effort of modelling the data well.

1. Split the data $\mathcal{D}$ into the training data $\mathcal{T}$ and test data $\mathcal{E}$.
2. Compute the reference trained model $P(X|\mathcal{T})$.
3. Compute the approximate model $\hat{P}(X|\mathcal{T})$.
4. Compute the reference maximum likelihood test model $\dot{P}(X|\mathcal{E})$.
5. The CV-value equals $\Pr\{D(\dot{P}(X)||P(X)) \geq D(\dot{P}(X)||\hat{P}(X))\}$.

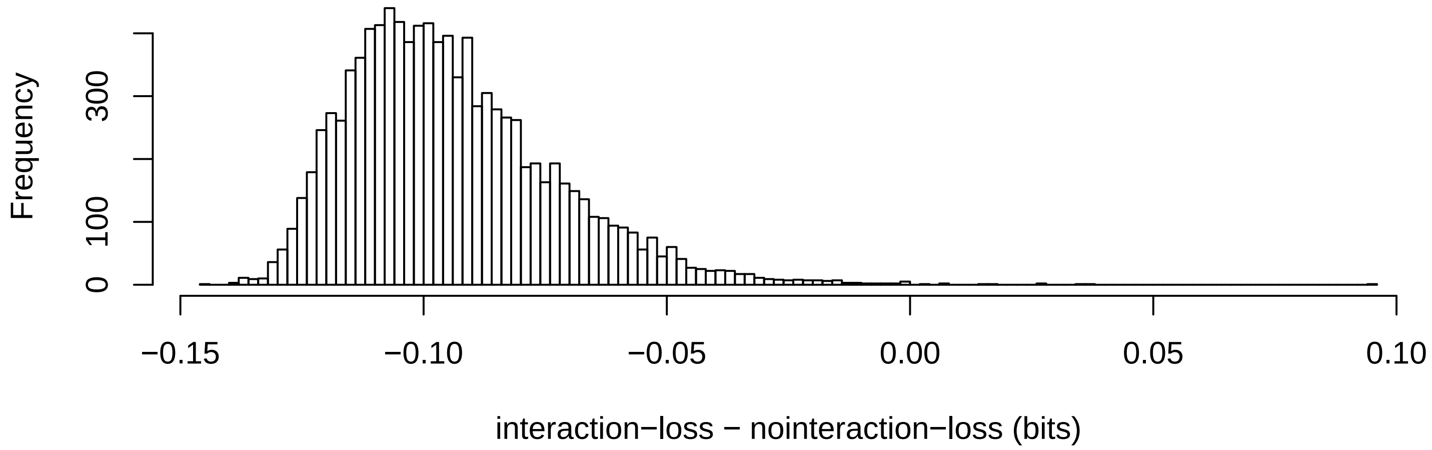
In generalization tasks, maximum-likelihood models are dangerous, as an outcome may have a zero occurrence count in the training data, but not in the test data. Bayesian priors can be used to remedy this problem of infinite divergences.

Good and Bad Attributes

A weak interaction (CV-value = 0.491, P-value = 0.690):



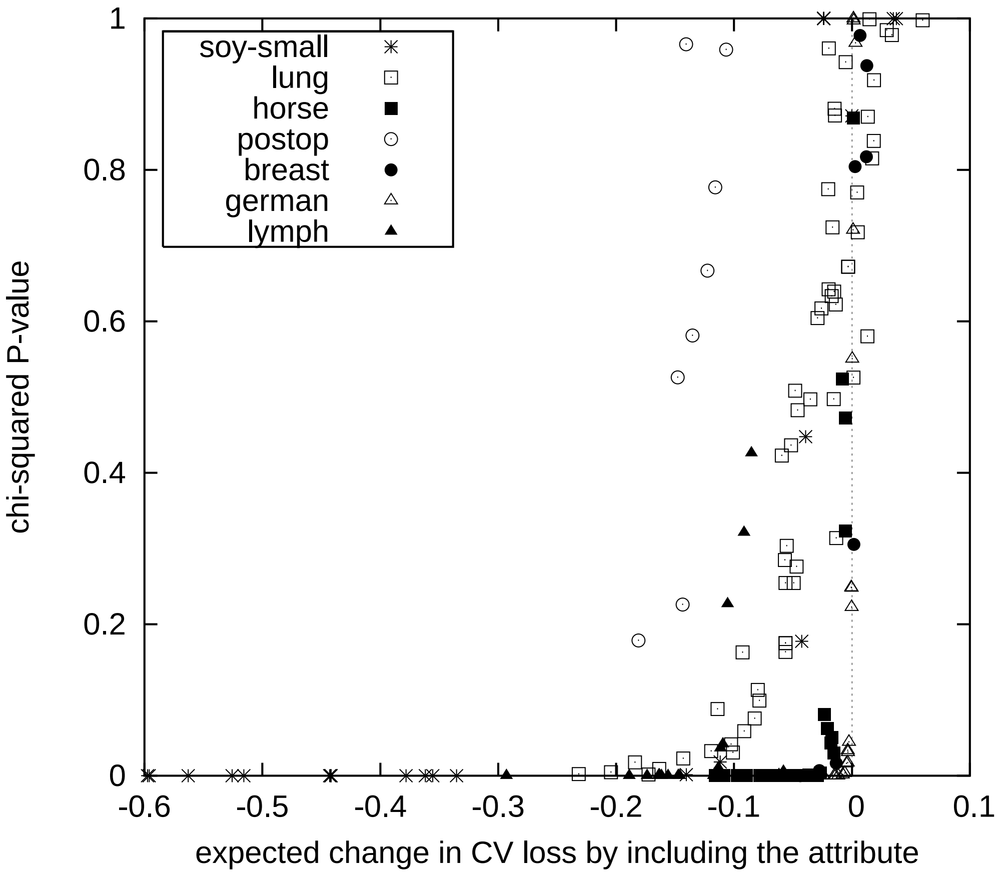
A strong interaction (CV-value = 0.001, P-value = 0.000):



The histograms illustrate the distribution of $D(\dot{P}(X)||\hat{P}(X)) - D(\dot{P}(X)||P(X))$

Expected Loss is Misleading!

If we only consider the expected change in loss computed on individual instances, it may happen that *highly insignificant interactions may reduce the expected cross-validated loss!* There is no justification for including an insignificant attribute into the model, except for its involvement in a significant higher-order interaction.



7. Summary

- Interactions are *irreducible dependencies*.
- *Interaction information* generalizes upon mutual information, and evaluates the irreducibility through Kirkwood superposition approximation (KSA) and the Kullback-Leibler divergence as the loss function.
- Negative interaction information is a consequence of KSA being *non-normalized*.

- Iterative scaling yields better models than KSA, but it is *not in closed form*.
- The *test-bootstrap* protocol explains how goodness-of-fit tests work.
- Kullback-Leibler divergence is both a loss function and a *statistic*, distributed with χ^2.
- *P-values* and *CV-values* evaluate the probability that using the interaction will result in an increase of loss.