Information-Theoretic Exploration and Evaluation of Models

Aleks Jakulin
Faculty of Computer and Information Science,
University of Ljubljana,
Tržaška cesta 25, SI-1001 Ljubljana, Slovenia.

Abstract

No information-theoretic quantity, such as entropy or Kullback-Leibler divergence, is meaningful without first assuming a probabilistic model. In Bayesian statistics, the model itself is uncertain, so the resulting information-theoretic quantities should also be treated as uncertain. Information theory provides a language for asking meaningful decision-theoretic questions about black-box probabilistic models, where the chosen utility function is log-likelihood. We show how general hypothesis testing can be developed from these conclusions, also handling the problem of multiple comparisons. Furthermore, we use mutual and interaction information to disentangle and visualize the structure inside black-box probabilistic models. On examples we show how misleading can non-generative models be about informativeness of attributes.

1 INTRODUCTION

The task of statistical modelling is to create reliable probabilistic models given the data and prior expectations about it. The prior expectations are both explicit Bayesian priors and implicit frequentist assumptions in the form of the choice of particular statistical models. Over the past years, automation of statistical modelling brought along models that are no longer easily presentable. This causes a dilemma: what are these models useful for? They may yield excellent performance, but how can we learn anything from them? How do we convey the meaning to a human analyst?

Shannon's information theory (Shannon, 1948) is a successful attempt to model communication and transmission of data. But it is important to note that there is nothing in information theory that would not be first based upon some generative probabilistic model. Even if the models used in information theory focus primarily on sequential and mainly on discrete data, information-theoretic tools are applicable to any joint probabilistic model, such as statistical models built for mainly real-valued data. Of course, probability mass functions differ from probability density functions, and some pitfalls must be avoided.

Information theory makes heavy use of concepts such as entropy and mutual information. We will interpret them through the usual statistical framework of models, model comparisons and loss functions. Namely, entropy, Kullback-Leibler divergence and mutual information are intrinsically decision-theoretic and not probability-theoretic concepts. They can be seen as a language for asking questions about the properties of models, and provide a good formalization of intuitive notions of relevance, dependence and complexity. Even if the underlying statistical model is a black box, information-theoretic notions can be used to ask questions about it, such as "How much would we lose by assuming the independence between these two attributes?" or "How much information about the outcome do we gain by this attribute?"

The first two sections on model loss and model comparison will present a few definitions which have been synthesized from works on information theory and machine learning. We will follow recent interpretations of entropy (Harremoës & Topsøe, 2001; Grünwald & Dawid, 2004) that view modelling as a game and entropy as a loss. Loss functions will be used as a foundation for model comparisons: loss is the sensible quantity to be used for model comparisons, rather than direct references to probability or parameter values. We will also follow up on the earlier work (Wolpert & Wolf, 1995; Hutter & Zaffalon, 2004) on probability distributions of information-theoretic quantities, noting that expected loss is an oversimplification in comparison to the view of loss as a stochastic quantity. The second part of the paper focuses on case studies.

2 MODEL LOSS

We begin by revising the basic terms, which are a mix of statistical and artificial intelligence terminology. An *instance* $\boldsymbol{i}$ corresponds to an event or an object described by a number of attributes. An *attribute* X is a unique property of instances that has a finite or an infinite *range* $\Re_X$ of mutually exclusive values. The value of the attribute X for the instance $\boldsymbol{i}$ is $x_{\boldsymbol{i}} \in \Re_X$. If there are several attributes, we may represent them together in an attribute vector $\boldsymbol{X} = [X_1, X_2, \ldots, X_M]$, and we refer to $\Re_{\boldsymbol{X}}$ as the attribute space. The joint probability density and mass functions (PDF, PMF) are models of co-appearance of individual attribute values in a randomly chosen instance.

Although entropy is often computed for an attribute or a set of them, Shannon did not define entropy for attributes, but for a joint *model* of the attributes, a particular joint probability mass function (PMF) P. Entropy should be seen as a characteristic of a model and not of an attribute or a data set. That is why expressing entropy as $H(A)$ is somewhat misleading; a more appropriate expression is $H(A|\phi, \mathcal{D})$, where $\mathcal{D}$ is the data and ϕ is the prior to the posterior predictive probabilistic model $P(A|\phi, \mathcal{D})$, the one actually used for computing the entropy. Although we will not always express entropy conditionally, the assumptions are always implicit: entropy is usually computed by assuming a maximum likelihood multinomial model.

Entropy H is defined for probability mass functions, not for probability density functions. For a multivariate joint PDF p modelling an attribute vector $\boldsymbol{X}$, a somewhat different concept of *differential entropy* h, also measured in bits, can be defined as (Cover & Thomas, 1991):

$$h(\boldsymbol{X}|p) \triangleq -\int_{\Re_{\boldsymbol{X}}} p(\boldsymbol{x}) \log_2 p(\boldsymbol{x}) d\boldsymbol{x} = E_p\{-\log_2 p(\boldsymbol{x})\} \tag{1}$$

The properties of differential entropy do not fully match those of ordinary entropy (Shannon, 1948). For example, differential entropy may be negative or even zero (e.g., $h(X|X \sim \mathcal{N}(\cdot, 1/\sqrt{2\pi e})) = 0$), and is sensitive to the choice of the coordinate system. Nonetheless, the magnitude of entropy and the sign of changes in entropy remain meaningful: the higher the entropy, the harder the predictions. Entropy should be understood as the expected loss of the model, given the model itself. Shannon entropy results from the choice of the logarithmic loss function. Other loss or utility functions may be employed and a corresponding generalized notion of entropy thus derived (Grünwald & Dawid, 2004), but its properties might not match those of Shannon entropy.

An analytical derivation of differential entropy has been made only for a few model families. Therefore, *empirical entropy* (Yeung, 2002), sometimes also referred to as sample entropy, often proves to be a useful approximation. If the data is a multiset of instances $\mathcal{D} \subset \Re_{\boldsymbol{X}}$, a probabilistic model $p(\boldsymbol{X}|\mathcal{D})$ can be learned from it. If the modelling is reliable, $\mathcal{D}$ can be understood as a representative random sample drawn from $\Re_{\boldsymbol{X}}$ using p. The approximation to h is the expected negative log-likelihood of a training instance given the model p:

$$\hat{h}(\boldsymbol{X}|p, \mathcal{D}) \triangleq -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \log_2 p(\boldsymbol{x}). \tag{2}$$

The resulting differential empirical entropy is the average negative log-likelihood of $\mathcal{D}$ given the model p. Observe that $1/|\mathcal{D}|$ is the probability of choosing a certain instance in $\mathcal{D}$. The resulting sum can then be understood as the expectation of entropy given a uniform probability distribution over the data: all instances in $\mathcal{D}$ have equal probability, and those outside are impossible.

3 MODEL COMPARISON

KL-divergence or relative entropy $D(P\|Q)$ (Kullback & Leibler, 1951) assesses the difference between two probability mass functions P and Q (or density functions p and q):

$$D(P\|Q) \triangleq \sum_{\boldsymbol{x} \in \Re_{\boldsymbol{X}}} P(\boldsymbol{x}) \log_2 \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} \tag{3}$$

$$D(p\|q) \triangleq \int_{\Re_{\boldsymbol{X}}} p(\boldsymbol{x}) \log_2 \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \tag{4}$$

KL-divergence is zero only when the two functions are equal. It is not a symmetric measure: P is the *reference* model, and the KL-divergence is the expected loss incurred by the *alternative* model Q when approximating P. We can understand empirical entropy through KL-divergence. If $U_{\mathcal{D}}$ is the uniform probability mass function on the data $\mathcal{D}$:

$$\hat{H}(\boldsymbol{X}|P, \mathcal{D}) = D(U_{\mathcal{D}}\|P) - H(U_{\mathcal{D}}), \tag{5}$$

$$U_{\mathcal{D}}(\boldsymbol{x}) \triangleq 1 - \frac{|\mathcal{D} \setminus \{\boldsymbol{x}\}|}{|\mathcal{D}|} \tag{6}$$

The same formula can be used to compute the differential empirical entropy of a PDF p, mixing probability mass and probability density functions, but ultimately yielding the same result as (2). If we interpret entropy as defined using KL-divergence, some problems of differential entropy, such as negativity, would be remedied with a better choice of the reference model U.

An important connection between entropy and KL-divergence appears when q is a marginalization of p:

$\int p(x,y)dx = q(y)$. In such a case, $D(p\|q) = h(p) - h(q)$. If q is a factorization of p, the KL-divergence can be expressed as a sum of entropies. Generally, the KL-divergence between p and any product of probability mass or density functions obtained by conditionalizing or marginalizing p is expressible by adding or subtracting entropies of p's marginals. For example, the divergence between $p(x,y,z)$ and $q(x,y,z) = p(x|z)p(y|z)p(z)$ is $h(x,z) + h(y,z) - h(z) - h(x,y,z)$. Mutual and conditional mutual information provide a short-hand notation, in this case: $D(p\|q) = I(x;y|z)$. Conditional and marginal entropies can be calculated through KL-divergence. Assuming $\boldsymbol{x} = [\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}]$, marginalizing over $\boldsymbol{c}$, the entropy of $\boldsymbol{a}$ conditionalized on $\boldsymbol{b}$ is $h(\boldsymbol{a}) - h(\boldsymbol{a}, \boldsymbol{b})$ or:

$$h(\boldsymbol{a}|\boldsymbol{b}) = \int_{\Re_{\boldsymbol{x}}} p(\boldsymbol{x}) \log_2 p(\boldsymbol{a}_{\boldsymbol{x}}|\boldsymbol{b}_{\boldsymbol{x}}) d\boldsymbol{x} = D(p(\boldsymbol{a}, \boldsymbol{b})\|p(\boldsymbol{b}))$$

Using conditional KL-divergence, it is possible to compare two conditional probability density functions, something particularly useful in supervised learning:

$$D(p(x|y)\|q(x|y)) = \iint p(x,y) \log_2 \frac{p(x|y)}{q(x|y)} dxdy \quad (7)$$

Observe, however, that conditional KL-divergence cannot be computed without a generative model of both x and y.

The main contribution of information theory in this context is the unified notation for model loss (H) and model comparisons (D). The expected negative log-likelihood is the agreed-upon model loss function (entropy), and the logarithm of the Bayes factor is the agreed-upon model difference function (KL-divergence). For empirical entropy and divergence, the expectation over the reference PDF/PMF is replaced with an expectation in a particular data set. All model comparisons are made between two probability functions: only models are compared and evaluated, not data and models.

4 DISTRIBUTIONS OF LOSS

It is customary to view all model comparisons as scalars. The implication of such an approach is that model comparisons are always crisp. In reality, however, one model is only sometimes better than another, and even the true model usually suffers decision-theoretic loss. For many decision-making purposes, the high cost of introducing any new model needs to be offset by the new model being consistently better than the prior or default one across all contingencies. These contingencies are often integrated out, but this yields a false sense of security as the actual loss is often greater than the expected one.

4.1 ENTROPY DISTRIBUTIONS

Empirical entropy is explicitly conditional on the sample, but 'ordinary' entropy is also conditional to a sample if the probabilistic model has been derived from data. There is a nuisance parameter, the hypothesis ϕ. It may be integrated out within the entropy computation, because entropy is computed using the predictive model. This results in a crisp estimate of entropy, as expected given the prior distribution $p(\phi)$ and the data $\mathcal{D}$:

$$h(\boldsymbol{x}|\mathcal{D}) = h\left(\boldsymbol{x}\middle|\mathcal{D}, \int p(\boldsymbol{x}, \phi|\mathcal{D}) d\phi\right) \quad (8)$$

Alternatively, we do not integrate it out, but model the probability distribution of entropy, as weighted by the probability of the posterior (Wolpert & Wolf, 1995):

$$\Pr\{h(\boldsymbol{x}|\mathcal{D}) \leq w\} = \int \mathbb{I}\{h(\boldsymbol{x}|\mathcal{D}, \phi)) \leq w\} p(\phi|\mathcal{D}) d\phi$$

Here, $\mathbb{I}$ is the indicator function, taking the value of 1 when the subscript condition is fulfilled and 0 otherwise. With this, we can investigate the variance of entropy estimates for an arbitrary probabilistic model.

We have not explained yet how to arrive at the distribution of KL-divergence. There are essentially two ways of approaching it. The first is to optimistically assume that both models, p and q share the same hypothesis ϕ (and the same evidence), so that one of the two models is nested within the other:

$$\Pr\{D_\phi(p\|q) \leq w\} = \int \mathbb{I}\{D(p_\phi\|q_\phi) \leq w\} p(\phi|\mathcal{D}) d\phi$$

This approach helps investigate the impact of a particular constraint or simplification relative to the original model p, and will be used for the examples in the present paper.

More generally, however, we should assume that the hypotheses (or evidence) are independently sampled:

$$\Pr\{D_{\phi|\rho}(p\|q) \leq w\} = \iint \mathbb{I}\{D(p_\phi\|q_\rho) \leq w\} \\ p(\phi|\mathcal{D}) p(\rho|\mathcal{D}) d\phi d\rho$$

An interesting special case of the second approach is the *self-divergence*, $D_{\phi|\phi}(p\|p)$; if we are uncertain about the hypothesis, this uncertainty should be reflected in the fact that there will generally be a difference between the two models obtained independently from the data.

4.2 MULTIPLE HYPOTHESIS TESTING

The notion of self-divergence is useful for hypothesis testing. As a rule of thumb, the expected self-divergence $E\{D_{\phi|\phi}(p\|p)\}$ of a reference model p should

be an order of magnitude lower than the expected loss of a model q based with respect to the reference model p, $E\{D_{\phi|\rho}(p\|q)\}$. If this is not the case, the reference model would be too complex given the data, and the variance of the estimate is not low enough to reliably estimate the bias. Both comparisons can be joined into a unique probability corresponding to a P-value for the comparison between the null hypothesis space $p(\phi|\mathcal{D})$ and the alternative hypothesis space $q(\rho|\mathcal{D})$, where KL-divergence is used instead of a test statistic. There is no need to draw samples from the model as with 'Bayesian' P-values because KL-divergence compares probabilistic models directly. The probability that $q(\rho|\mathcal{D})$ is worse than an independent $p(\phi'|\mathcal{D})$ is:

$$\iiint \mathbb{I}\{D(p_\phi\|q_\rho) \leq D(p_\phi\|p_{\phi'})\}p(\phi|\mathcal{D})p(\rho|\mathcal{D})p(\phi'|\mathcal{D})d\phi\rho\phi'$$

Multiple testing is trivial in this framework: all that needs to be changed is the index function. It would be an oversimplification to assume independence between hypotheses, as by Bonferroni correction, if these hypotheses relate to the same data and potentially to the same variables. In any case, correct multiple testing becomes increasingly difficult with a large number of variables without making strong prior assumptions. For example, the test for the conjunction of statements "$\boldsymbol{x}$ is independent of $\boldsymbol{y}$" and "$\boldsymbol{x}$ is independent of $\boldsymbol{z}$" would have to be based on models $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\mathcal{D})$ conditioned on priors ϕ and ρ with the following indicator function:

$$D(p(\boldsymbol{x}, \boldsymbol{y}|\phi)\|p(\boldsymbol{x}|\rho)p(\boldsymbol{y}|\rho)) \leq D(p(\boldsymbol{x}, \boldsymbol{y}|\phi)\|p(\boldsymbol{x}, \boldsymbol{y}|\phi'))\wedge$$
$$D(p(\boldsymbol{x}, \boldsymbol{z}|\phi)\|p(\boldsymbol{x}|\rho)p(\boldsymbol{z}|\rho)) \leq D(p(\boldsymbol{x}, \boldsymbol{z}|\phi)\|p(\boldsymbol{x}, \boldsymbol{z}|\phi'))$$

The entropy and KL-divergence distributions are not symmetric, so it is often preferable to work with percentiles. The distributions can be used to obtain the entropy confidence intervals, e.g. between the 97.5th and 2.5th percentiles. If a single-valued estimate is needed, the practical worst-case loss would be the loss at the 99th percentile, analogous to value-at-risk decision making. Expected loss is overly optimistic and can cause gambler's ruin with a non-negligible probability, while worst-case loss is usually infinite.

5 EXAMPLES

5.1 CORRELATION

Correlation analysis is one of the most frequently used tools of classical statistics, yet it is often missing from textbooks on Bayesian statistics. We will now investigate correlation analysis as a special case of model comparison. Correlation will be quantified decision-theoretically through bits of information gained by allowing a rotation of the input attribute space. As the model comparison is based on two random models, the information gained is a random variable, which can be subjected to generic model comparisons of Sect. 4. No special purpose 'correlation coefficients' or special purpose tests of correlation are necessary: we simply compare different models.

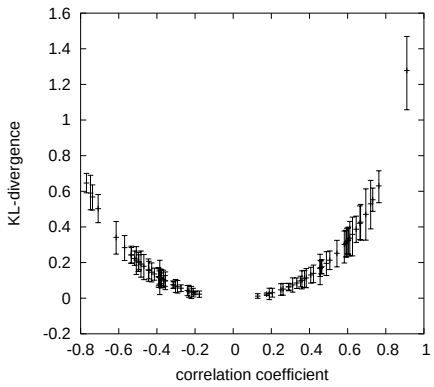

Figure 1: We may express correlation using KL-divergence instead of the correlation coefficient. While retaining monotonicity, the scale of KL-divergence is more intuitive than that of the correlation coefficient, noting the rule of thumb that correlation coefficients with the absolute value lower than 0.3 are not interesting. In this figure we plot the importance of correlation for all pairs of attributes in the Boston housing data. The wide confidence interval on the extreme right should raise suspicion: the high correlation for that particular pair of attributes (*property tax* and *highways*) is merely due to a few high property tax outliers. The next two correlations, (*nitric oxides* with *employment distance* and *non-retail acres*) are more meaningful and more stable.

A simple reference model p that allows correlation is a multivariate normal distribution. The d-dimensional attribute vector $\boldsymbol{x} = [x_1, \ldots, x_d]$ is treated as a single multi-dimensional attribute. On the other hand, the alternative model q models each attribute independently.

$$p: \quad \boldsymbol{x} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{9}$$

$$q: \quad \boldsymbol{x} \sim \prod_i^d \text{Normal}(\mu_i, \sigma_i) \tag{10}$$

This scheme is not limited to two dimensions, so correlations involving an arbitrary number of attributes can be investigated easily. Furthermore, it is not necessary for the covariance matrix $\boldsymbol{\Sigma}$ to be orthogonal: some additional information can be gained by this. Fig. 1 demonstrates the relationship between the KL-divergence and the correlation coefficient ρ: $D(p\|q) = -\frac{1}{2}\log_2(1-\rho^2)$ (Billinger, 2004).

5.2 INTERACTION

There are many interpretations of what an interaction is. Usually, interaction is thought to be a term that combines multiple attributes. However, we will instead interpret an interaction decision-theoretically as the benefit gained from a model of multiple attributes in comparison to a fusion of models based on subsets of attributes. Thus, a k-way interaction among k groups of attributes $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k\}$ is the reduction in loss achievable by using the joint model of k attributes $p(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k)$ in comparison to its part-to-whole approximation reconstructed solely from p's marginals $\mathcal{M} = \{\int p(\mathcal{A})d\mathcal{A}_i;\ \mathcal{A}_i \in \mathcal{A}\}$ (Jakulin & Bratko, 2004). A comparison of the reference model p and its part-to-whole approximation is effectively a test of a k-way interaction within $\mathcal{A}$ in p.

The task of making a general part-to-whole approximation is thought to be a difficult modelling problem, usually approached through maximum entropy modelling, iteratively adjusting the joint model to be consistent with the given marginal models constraining it (Nemenman, 2004), and maximizing its entropy. However, it is also possible to employ the notion of McGill's interaction information and the corresponding Kirkwood superposition approximation where the part-to-whole approximation emerges in closed form composed only from a product of marginalizations of p (Jakulin & Bratko, 2004). The KL-divergence between p and its part-to-whole Kirkwood superposition approximation matches the definition of *interaction information* for a set of attributes $\mathcal{A}$:

$$\hat{\imath}(\mathcal{A}|p,\mathcal{D}) \triangleq -\sum_{\mathcal{X}\subseteq\mathcal{A}} (-1)^{|\mathcal{A}\setminus\mathcal{X}|}\hat{h}(\mathcal{X}|p,\mathcal{D}) \qquad (11)$$

For the case of three variables, interaction information corresponds to $I(A;B;C) = I(A,B;C) - I(A;C) - I(B;C)$. This can be interpreted as the difference between the true informativeness of attributes A and B about the quantity of interest C on one hand, and the sum of individual contributions of A and B to information about C. If $I(A;B;C)$ is distinctly positive, we can say that there is a synergy between A and B when predicting C. If it is distinctly negative, there is a redundancy where both A and B provide partially the same information about C.

Interaction information has proven to be a considerably better predictor of validity of the NBC assumption in classification tasks than conditional mutual information $I(A;B|C)$. This can be apparent from the identity, remembering (7):

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

The right-hand model closely resembles an non-normalized naïve Bayesian classifier (NBC). This non-normalization is what yields a negative interaction information, and $I(A;B;C)$ should really be seen as an approximate model comparison (but with many other convenient properties). Conditional mutual information tends to overestimate the deviation, as it is derived from a joint model comparison, and not a conditional one. Conditional independence relations can also be represented in the part-to-whole context. For example, if the part-to-whole approximation created from $\mathcal{M} = \{P(A,B), P(A,C)\}$ is indistinguishable from $P(A,B,C)$, then B and C are conditionally independent given A. The part-to-whole approximation for such marginals can be obtained using the chain rule in closed form by conditioning on A.

Before applying interaction analysis, we first need the underlying probabilistic model. It is not necessary to have a single global model: through local analysis we build a separate model for each subset of attributes under investigation, as the global model would match the local one if the attributes outside the focus were marginalized away. Really, marginalization can be performed both on the data or on the model. Mixture models will be our hypothesis space. We will make the assumption of *local independence*, so that the latent attribute Z will account for all the dependence between attributes $\boldsymbol{X} = [X_1, X_2, \ldots, X_d]$:

$$p(\boldsymbol{X}|Z) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{d} p(X_i|\phi_{k,i}) \qquad (12)$$

Each individual value of Z can be interpreted both as a *component*, a probabilistic prototype, a *cluster* a set of instances that correspond to the prototype (to some extent), or as an *axis* or dimension of a vector space where the instances can be represented as points. The choice of the functions in the mixture depends on the type of the attribute. Most implementations are based on normal or Gaussian mixtures, which work only for continuous attributes. The MULTIMIX program (Hunt & Jorgensen, 1999), however, handles both continuous and discrete attributes simultaneously, adopting the multinomial distribution for any discrete attribute and the normal distribution for any continuous attribute.

To demonstrate the analysis of mixture models using interaction information, we analyzed two UCI regression data sets, 'imports-85' and 'Boston housing'. For each potential interaction, a five-component joint mixture model was built. Because both data sets are regression problems, the outcome was always included in the model. The three kinds of models were: (1) the outcome alone, (2) each attribute with the outcome, and (3) each pair of attributes with the outcome. The

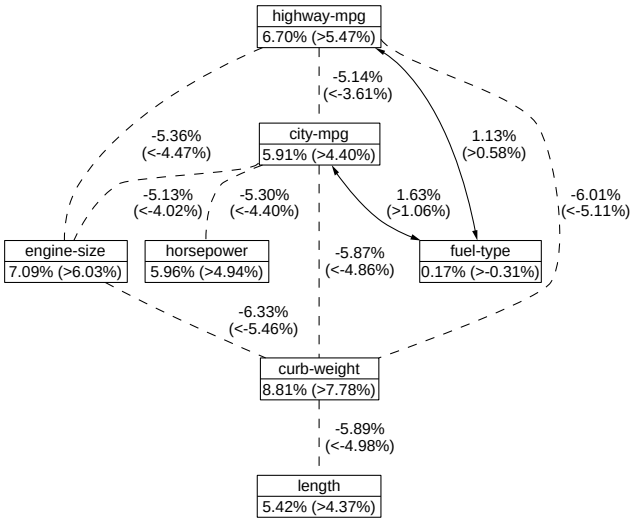

Figure 2: The interaction graph identifies the strongest 2-way and 3-way interactions in the 'imports-85' data set.

interaction information for each of these models was estimated along with its 95% confidence interval. For performance reasons we approximated the distribution of loss by computing the KL-divergence for each instance, and using the bootstrap replications over instances to compute the percentiles.

The models whose deviation from the part-to-whole approximation was the largest, either with positive or negative interaction information, were displayed graphically: the attributes entangled with the outcome in a 2-way interaction appear as nodes, and the pairs of attributes entangled with the outcome in a 3-way interaction appear as links connecting the nodes. The interaction information was expressed as a percentage of the outcome entropy alone. The percentages are not always sensible for probability density functions, but with care can nevertheless be more interpretable than bits of information.

Figure 2 shows the interaction graph on the 'imports-85' regression problem. The outcome being predicted is the value of the car, while other attributes describe the car's properties. The numbers below each attribute indicate the proportion of label entropy the attribute eliminates with a 97.5% bottom bound. For example, *highway mpg* alone eliminates 6.7% of uncertainty about the price on average, but in 97.5% of cases more than 5.5%. The best individual attribute is the weight of the car, eliminating more than 8.8% of outcome uncertainty. *Fuel type* may appear to be a useless attribute on its own, eliminating only 0.2% of outcome entropy, but there is a positive interaction or a synergy between fuel type and fuel consumption on the highway, eliminating an additional 1.1% of label entropy; the fuel consumption should be viewed in the context of whether the vehicle consumes gasoline or diesel fuel. Dashed edges indicate negative interactions or redundancies, where two attributes provide partly the same information about the label. Should we consider the fuel consumption both on highways and in the city, the total amount of label entropy eliminated is $6.7 + 5.9 - 5.1$ percent, the 5.1% accounting for their overlap. Due to the imprecision of empirical entropy and unsupervised modelling criteria, seeming illogicalities may appear: the length of the automobile is hurting the predictions of the car's price in combination the car's weight because the limited complexity of the model is spent for increasing the likelihood of the joint model rather than optimizing the prediction of the outcome.

Figure 3 illustrates the result of interaction analysis of the 'Boston housing' data set. The outcome of interest is the median value of apartment in a certain area, as predicted by various properties of the area, such as unemployment, crime rate, pollution, etc. The most informative attribute is the proportion of lower status population. In the context of this attribute, *non-retail acres* becomes almost totally uninformative ($7.99 - 7.93 = 0.06$). Another useful attribute is *crime-rate*, which subsumes most of the information provided by *prior-1940* and *employment-dist*. Furthermore, a strong negative interaction between *pupil-teacher* and *nitric-oxides* must be noted. Although most negative interactions are due to correlations between attributes, these two are themselves not highly correlated, and the negative interaction is nonlinear in character. At low levels of pollution, the housing value is mostly independent of pollution given the pupil-teacher ratio. On the other hand, at higher levels of pollution, the pupil-teacher ratio does not vary. Using the above interaction graph, it is also possible to understand why *non-retail acres* and *prior 1940* prove to be insignificant (with P-values of 0.74 and 0.96, respectively) in a multiple regression model (R Development Core Team, 2004), even if they are significant on their own:

	Estimate	Std.Error	t-val	Pr(>|t|)	
(Intercept)	3.646e+01	5.103e+00	7.144	3.28e-12	***
crime.rate	-1.080e-01	3.286e-02	-3.287	0.001087	**
zoned.lots	4.642e-02	1.373e-02	3.382	0.000778	***
non.retail.acres	2.056e-02	6.150e-02	0.334	0.738288	
Charles.River	2.687e+00	8.616e-01	3.118	0.001925	**
nitric.oxides	-1.777e+01	3.820e+00	-4.651	4.25e-06	***
rooms	3.810e+00	4.179e-01	9.116	< 2e-16	***
prior.1940	6.922e-04	1.321e-02	0.052	0.958230	
employment.dist	-1.476e+00	1.995e-01	-7.398	6.01e-13	***
highways	3.060e-01	6.635e-02	4.613	5.07e-06	***
property.tax	-1.233e-02	3.761e-03	-3.280	0.001112	**
pupil.teacher	-9.527e-01	1.308e-01	-7.283	1.31e-12	***
B	9.312e-03	2.686e-03	3.467	0.000573	***
low.status	-5.248e-01	5.072e-02	-10.347	< 2e-16	***

These attributes are not irrelevant, they merely become insignificant in the context of other attributes,

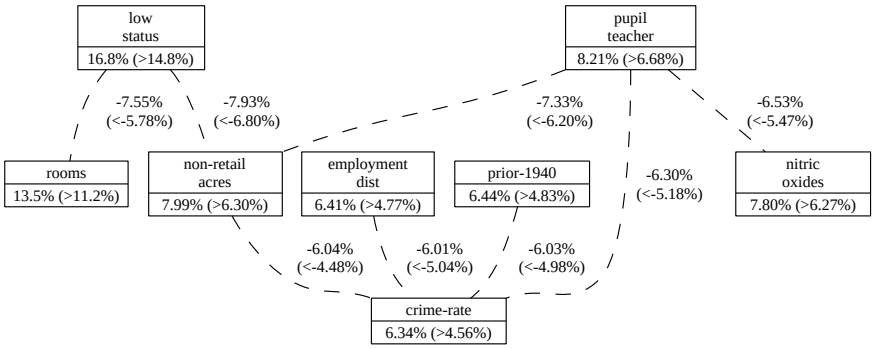

Figure 3: The strongest two-way and three-way interactions involving the label in the 'Boston housing' data set.

such as *low status*. Of course, deciding which attribute should get the credit for predicting the outcome is often arbitrary: we may greedily credit just the best attribute, or we may be egalitarian in distributing the information credit among them all.

5.3 STRUCTURE

Another useful discovery that can be made about the data is evidence for multiple groups in data. Generally, the decision-theoretic value of structure is the reduction in entropy achieved by using K instead of K' components in a finite mixture model, $K > K'$. Structure allows a relatively simple model to capture complex non-linear relationships in data, not just multimodality. Through local analysis, we may investigate the structure aspect in small subsets of attributes, seeking useful patterns and trying to localize the complexity. The results of such analysis are illustrated in Fig. 4 for the 'Boston housing' data set, using these two models:

$$p: \quad \boldsymbol{x} \sim \sum_{k=1}^{5} \pi_k \prod_{i}^{d} \text{Normal}(\mu_{k,i}, \sigma_{k,i}) \tag{13}$$

$$\boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\lambda}, 1), \ \sum_{k} \lambda_k = 1 \tag{14}$$

$$q: \quad \boldsymbol{x} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{15}$$

The purpose of such analysis is to discover interesting projections of the data, thus guiding exploratory data analysis and data mining. Gain of information through structure or correlation is what makes a projection interesting.

6 DISCUSSION

Most problems in estimating entropy of data can be reduced to finding a good probabilistic model of the data. Apart from having appealing properties (some of which are not retained by empirical entropy), entropy can be seen as a prototypical loss function which measures the quality of a particular model, and, as information, the worth of changes to the model.

We can also compare different philosophies of statistics. While the Fisherian self-divergence used for significance testing arises from the correspondence of multiple finite samples (all but one hypothetical) of the same size to the same given (maximum likelihood) model, the Bayesian self-divergence we describe in Sect. 4 is due to the consistency of multiple models with the same given finite sample. In Fisherian significance testing, the the test statistic assesses the agreement between a sample and a model, whereas in Bayesian significance testing the divergence functions quantify the fit between the two models.

We have shown how flexible applications of model comparison can be used to discover correlations and structure, even if the modelling is local, for subsets of attributes. However, both structure and correlation can be seen just as special cases of interaction: the undesirability of factorizing a probabilistic model. Thinking in terms of interactions is important for several reasons:

- Interactions of different kinds are often supported by the data. A flexible hypothesis space, such as that of log-linear or mixture models, will be able to capture them, while other approaches to modelling will often assume them away.
- High-order interactions are intrinsically difficult to model due to a large number of parameters and the resulting model uncertainty. Therefore it is often necessary to assume restrictions on certain types of interactions.
- The presence or absence of an interaction between a subset of attributes is highly interpretable and informative to a human analyst, along with a projection of the data to just that subset. This kind of analysis can be performed either for a global

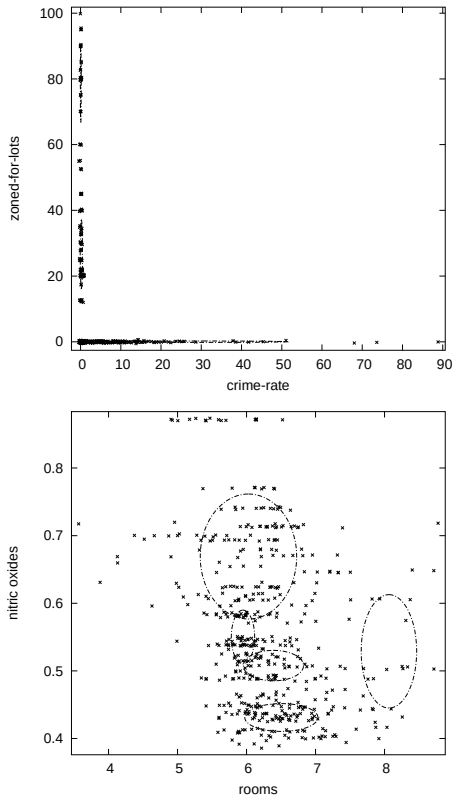

Figure 4: For the 'Boston housing' data set, the scatter plot on the top illustrates the nonlinear dependence between *crime rate* and *zoned for lots*, which has the highest amount of structure among all attribute pairs. On the other hand, structure is not of considerable utility to the model of *nitric oxides* and *rooms* (bottom). Axis-aligned ellipses depict the circumference of each component at one standard deviation in the reference mixture model.

black-box probabilistic model, or for local models.

- Conditional modelling may provide misleading insights into the informativeness of individual attributes, due to attribute redundancies and synergies. It is better to compare models than to examine model parameters.

We will conclude with a motivational example. Consider this example of a greedily built regression model for car prices:

```
             Estimate  Std.Error   t-val  Pr(>|t|)
(Intercept) -32254.698  17385.307  -1.855   0.0651 .
curb.weight     13.126      1.406   9.333   <2e-16 ***
width          753.987    313.931   2.402   0.0173 *
height        -316.178    148.979  -2.122   0.0351 *
length        -119.198     64.586  -1.846   0.0665 .
```

We could not claim that width, height and length of an automobile are uninformative about its price, in spite of the model. This pitfall inherent to conditional models is avoided by constructing joint models, performing model comparisons, and by using information-theoretic examination of interactions in the models. This way, looking at Fig. 2, we would observe that *length* gives us very little additional information about the car price once we already know *curb.weight*, but in case the weight is not known, length alone is nevertheless quite a useful attribute.

References

Billinger, D. R. (2004). Some data analyses using mutual information. *Brazilian J. Probability and Statistics*. to appear.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley, New York.

Grünwald, P. D., & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, *32*(4).

Harremoës, P., & Topsøe, F. (2001). Maximum entropy fundamentals. *Entropy*, *3*, 191–226.

Hunt, L., & Jorgensen, M. (1999). Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, *41*(2).

Hutter, M., & Zaffalon, M. (2004). Distribution of mutual information from complete and incomplete data. *Computational Statistics & Data Analysis*. to appear.

Jakulin, A., & Bratko, I. (2004). Testing the significance of attribute interactions. In Greiner, R., & Schuurmans, D. (Eds.), *Proc. of 21st International Conference on Machine Learning (ICML)*, pp. 409–416 Banff, Alberta, Canada.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Ann. Math. Stat.*, *22*, 76–86.

Nemenman, I. (2004). Information theory, multivariate dependence, and genetic network inference. Tech. rep. NSF-KITP-04-54, KITP, UCSB.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Wolpert, D. H., & Wolf, D. R. (1995). Estimating functions of distributions from a finite set of samples. *Physical Review E*, *52*(6), 6841–6854.

Yeung, R. W. (2002). *A First Course in Information Theory*. Kluwer Academic/Plenum Publishers, New York.