DLR

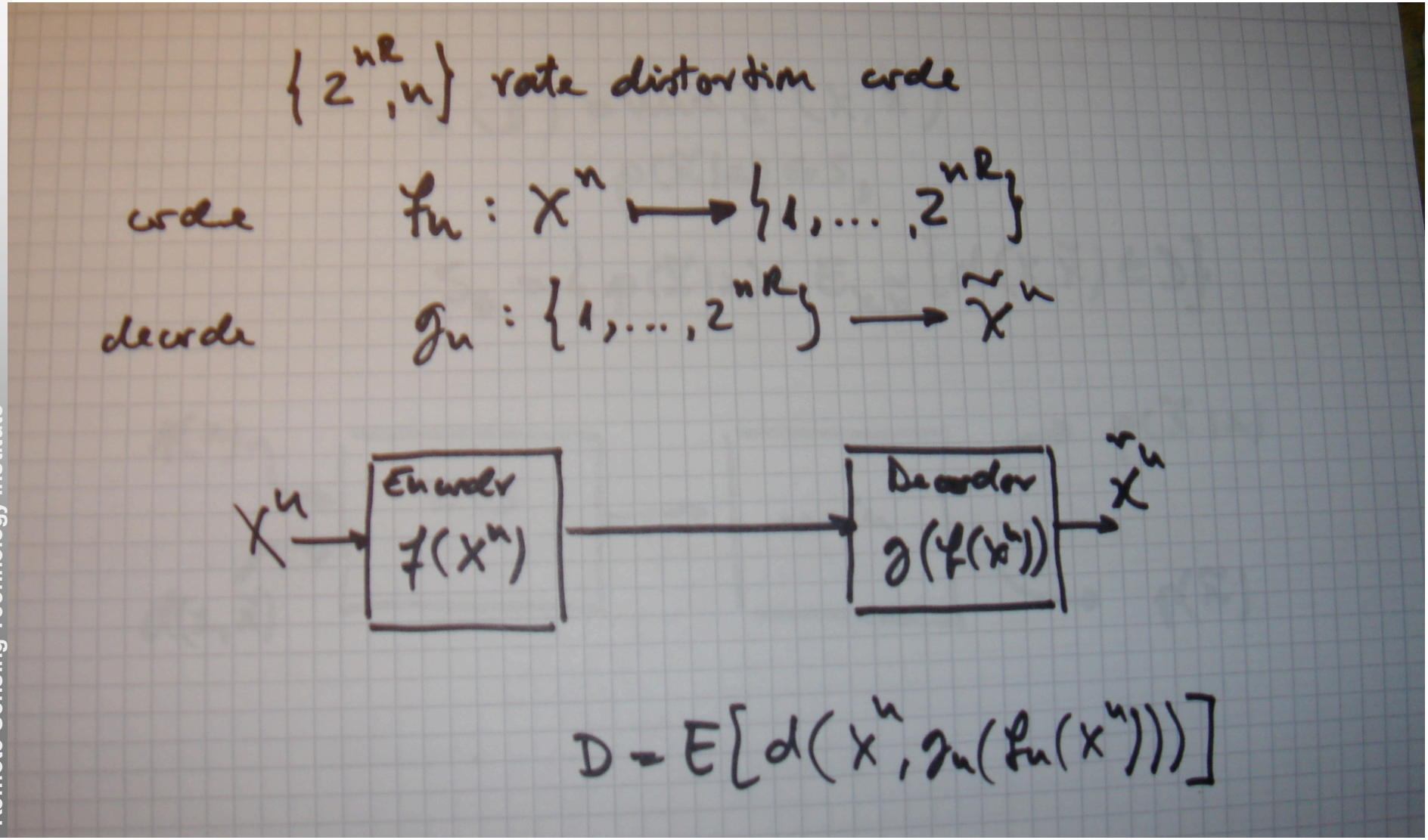
Information Theory based Inference: applications to image mining and semantic coding

INFORMATION BOTTLENECK (cont.)

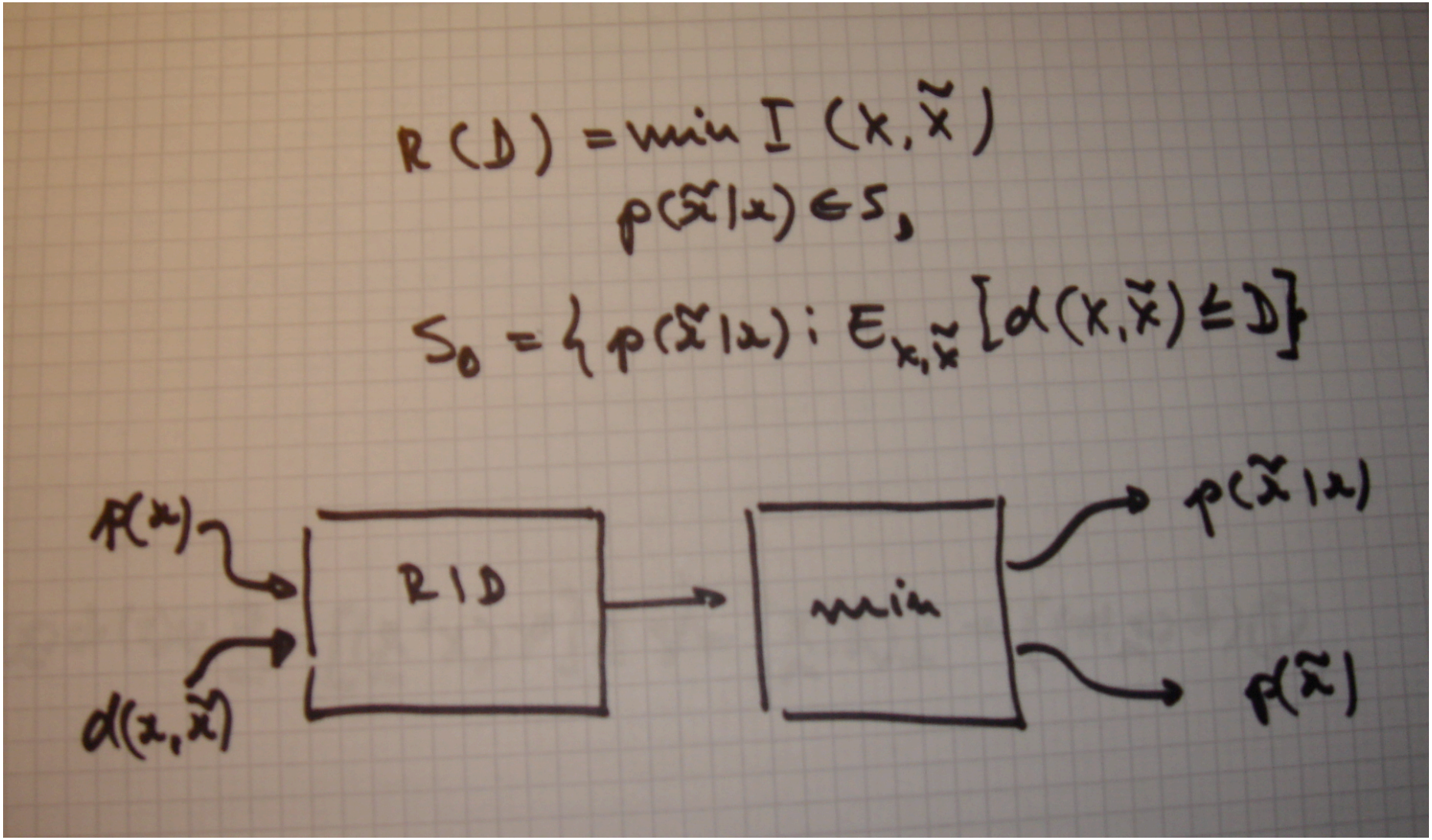
Mihai Datcu

Remote Sensing Technology Institute

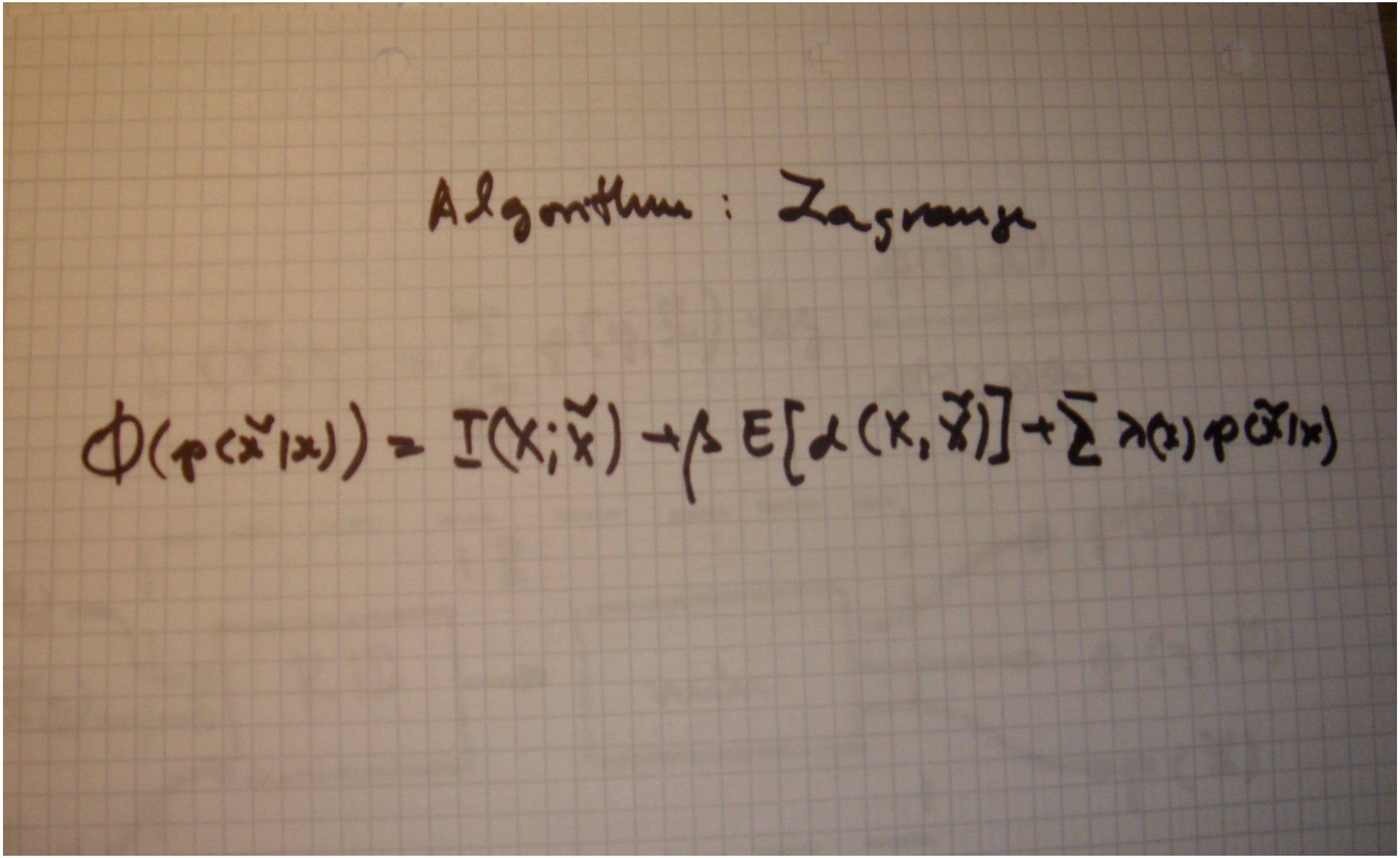
Coding


$\{2^{nR}, n\}$ rate distortion code
code $f_n : X^n \longrightarrow \{1, \dots, 2^{nR}\}$
decode $g_n : \{1, \dots, 2^{nR}\} \longrightarrow \tilde{X}^n$
X^n
Encoder
$f(X^n)$
Decoder
$g(f(X^n))$
$\tilde{X}^n$
$D = E[d(X^n, g_n(f_n(X^n)))]$

Rate Distortion: the inference



Computation



IB overview

Information Bottleneck extends Rate Distortion theory to relevant information extraction

Relevant information is the information in a pattern needed to predict semantics or help recognition

Rate Distortion theory is applied to maximize the information retained for a particular length description

Questions

How to extract good representations of the relevant information contained in a large set of features?

What information is relevant?

Information Bottleneck provides:
- Selected Features
- Efficient Representation

Definitions

Relative Entropy (Kullback Liebler distance) is a measure of the inefficiency of assuming distribution q when the true distribution is p

Mutual Information is the relative entropy between the joint distribution and the product distribution

Interpretation: reduction in uncertainty of X due to knowledge of Y

Concept

Find how well we can represent a r.v. using a compressed representation

Quality is defined as minimizing the rate and maximizing the information captured about the relevance variable

The information in the compressed representation is the mutual information

Concept

The Information Bottleneck is an extension of the Rate Distortion Theorem using a supervised definition of relevance

Concept

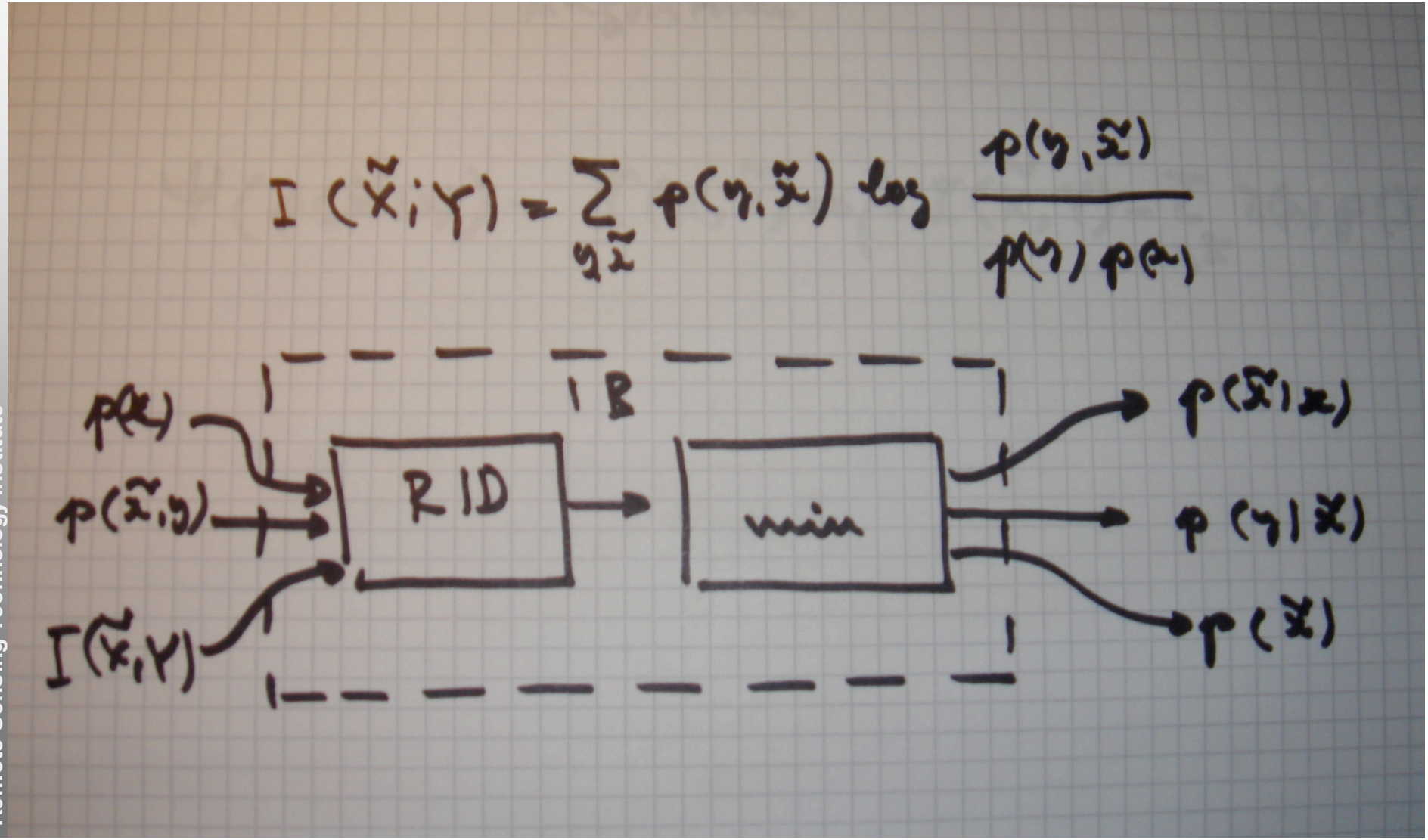
Information Bottleneck method applies Rate Distortion theory to supervised information extraction

Relevant information is defined as the information useful for predicting

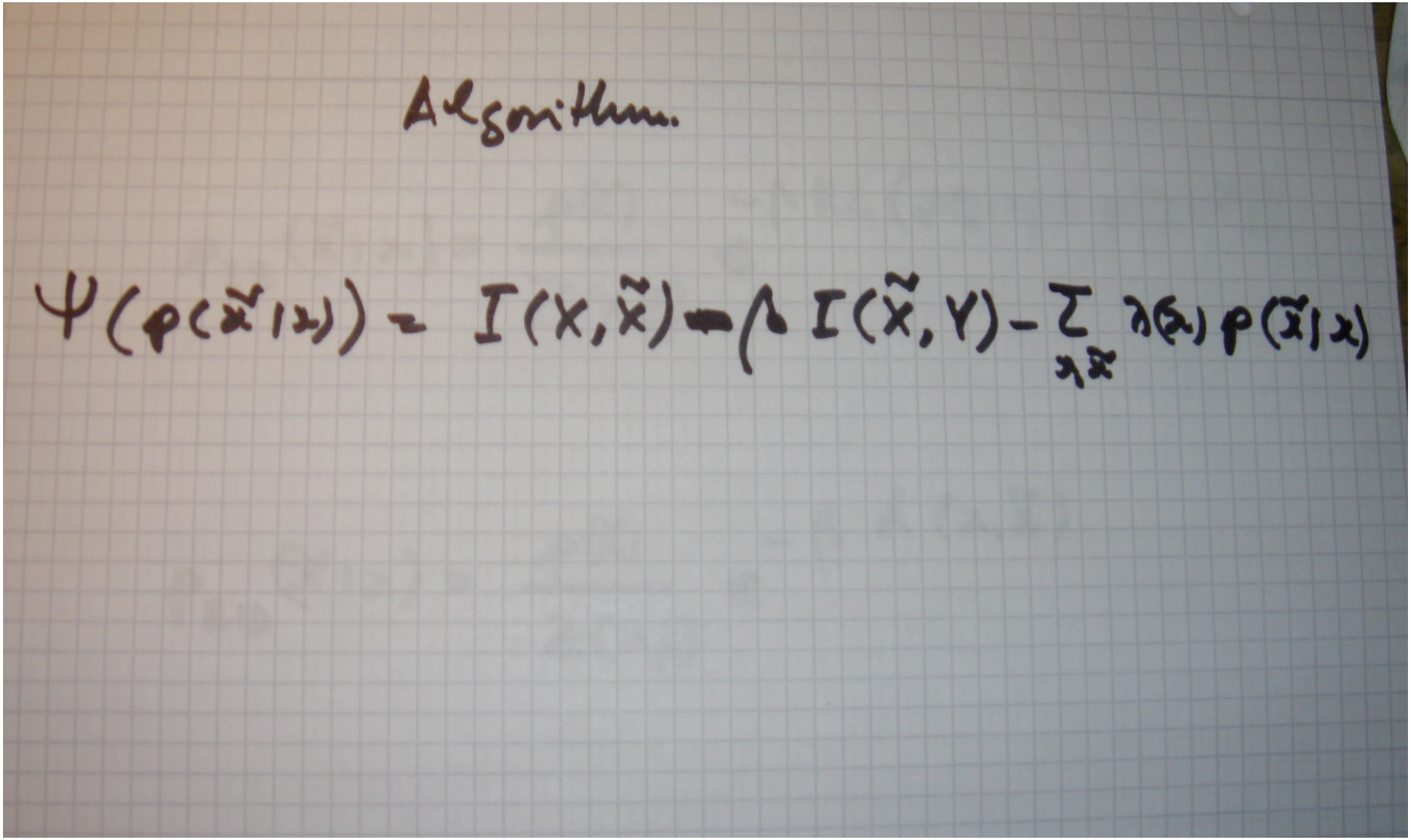
An iterative minimization algorithm is applied to find the partitioning

The solution is equivalent to using the KL distance

Information Bottleneck: the inference



Computation



R/D vs. IB

$$p_{IB}(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} e^{-\beta KL(p(y|x) \| p(y|\tilde{x}))}$$

$$p_{RD}(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} e^{-\beta d(x,\tilde{x})}$$

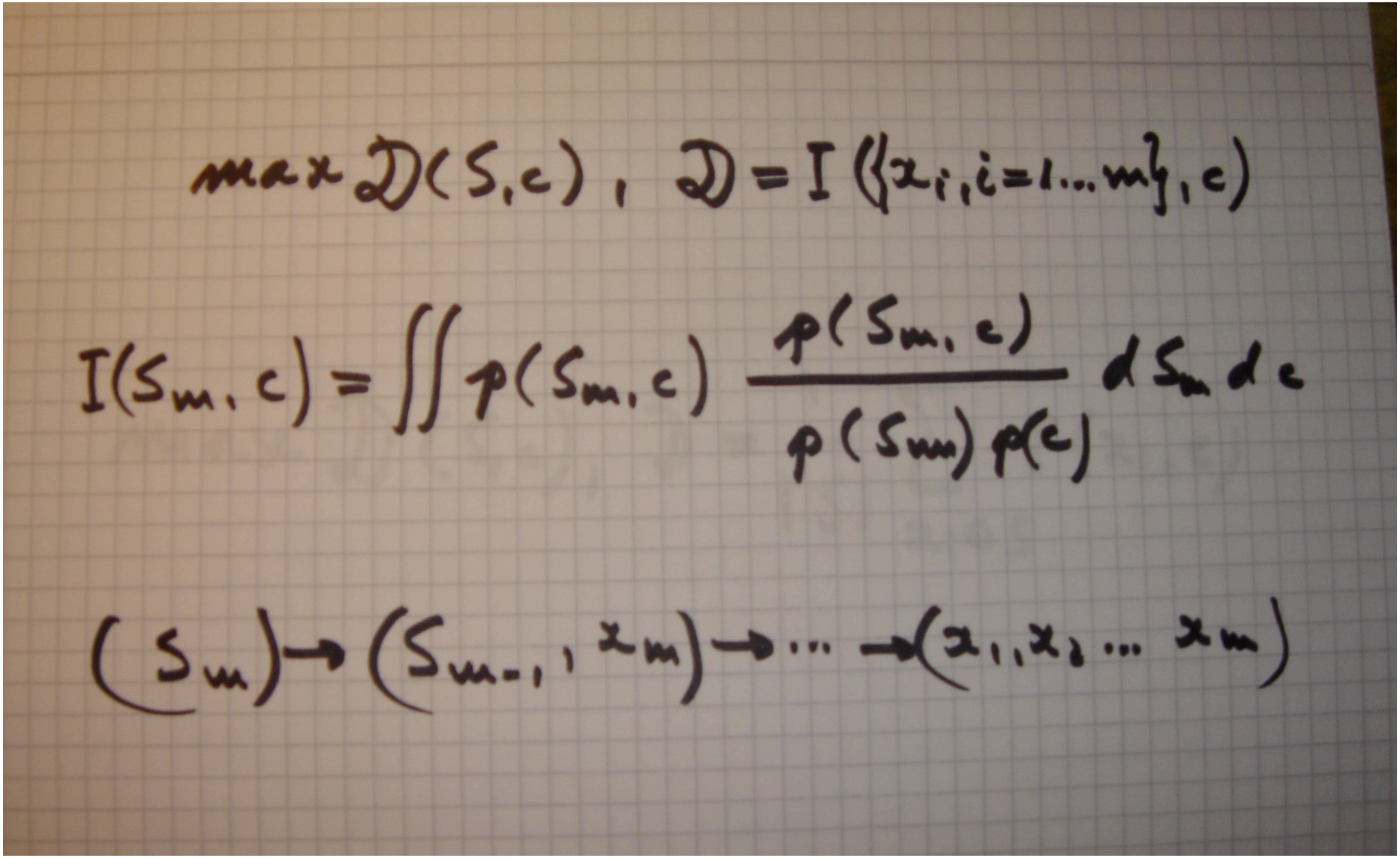
Max-Dependency

Optimal characterization condition often means the minimal classification error

In an unsupervised situation where the classifiers are not specified, minimal error usually requires the maximal statistical dependency of the target class c on the data distribution in the signal subspace

This scheme is maximal dependency (Max-Dependency).

Max-Dependency



Max-Relevance

To realize Max-Dependency is maximal relevance (Max-Relevance) feature selection: selecting the features with the highest relevance to the target class c

Relevance is usually characterized in terms of correlation or mutual information

Mutual information is the measure to define dependency of variables

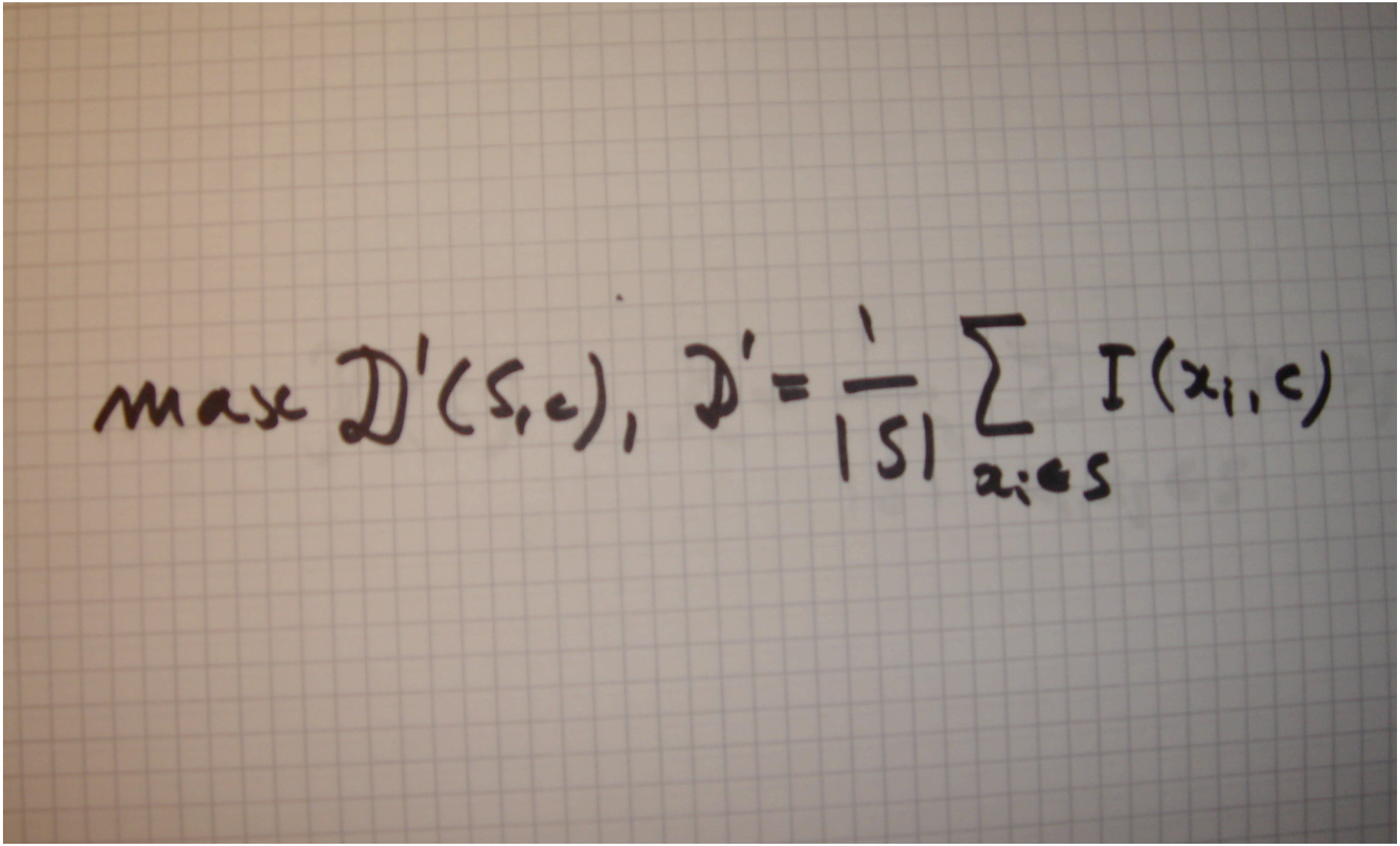
Max-Relevance

Max-Relevance, the selected features are required individually to have the largest mutual information with the target class c

That is reflecting the largest dependency on the target class

In terms of sequential search: the m best individual features, i.e., the top m features in the descent ordering of are often selected as the m features.

Max-Relevance



“the m best features are not the best m features”

Feature selection:

it has been recognized that the combinations of individually good features do not necessarily lead to good classification Performance

“the m best features are not the best m features”

Max Dependecy

incremental search scheme is to add one feature at one time

given the set with m 1 features, Sm 1, the mth feature can be determined as the one that contributes to the largest increase of which it takes

Max-dependency

Often hard to get an accurate estimation for multivariate density because in the high-dimensional space:
- the number ofsamples is often insufficient
- the multivariate density estimation often involves computing the inverse of the high-dimensional covariance matrix, (illposed problem)

Slow computational speed.
Not appropriate for applications where the aim is to achieve high classification accuracy with a reasonably compact set of features.

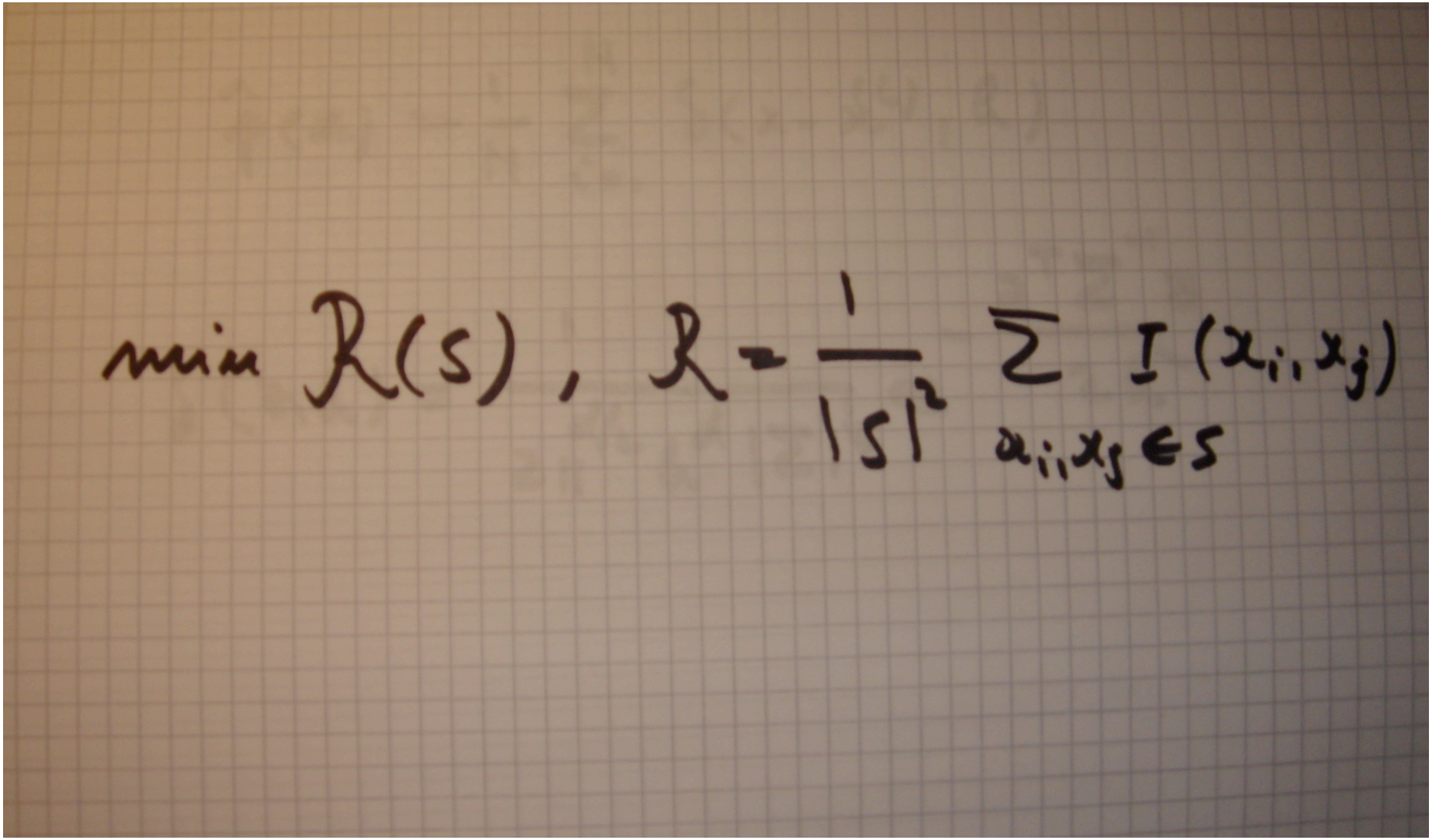
Max-Relevance

Search features which approximates Max_Dependency with the mean value of all mutual information values between individual feature xi and class c

Min-Redundancy

Add condition to select mutually exclusive features

Min-Redundancy



Parzen approximation

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)}, h)$$

$$\delta(z, h) = \frac{1}{2\pi^{d/2} h^d |\Sigma|^{\frac{1}{2}}} e^{-\frac{a^T \Sigma^{-1} a}{2h^2}}$$