BAYESIAN METHODS

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma
(Ed. CU013, st.211 0649914367)
Andrea.giansanti@uniroma1.it

CB_23_24 L n. 28, 29 and 30, Roma 4,5 and 7 dec 2023 .

SAPIENZA UNIVERSITÀ DI ROMA
Dipartimento di Management
DIPARTIMENTO DI FISICA

OUTLINE OF LECTURE CB_23_24_L28

relevance of Bayes' theorem in the analysis of sequences
generative probabilistic models
Markov order 0 models (urn models)

A bayesian classifier of disordered proteins (a critique of, look at the priors) (Bulashevska2008)

multinomial classification
look at the priors

The relevance of Bayes' theorem: see DILL & BROMBERG: EXAMPLE1.11 …BIOINFORMATIC CONTEXT

EXAMPLE 1.11 Applying Bayes' rule: Predicting protein properties. *Bayes' rule*, a combination of Equations (1.11) and (1.15), can help you compute hard-to-get probabilities from ones that are easier to get. Here's a toy example. Let's figure out a protein's structure from its amino acid sequence. From modern genomics, it is easy to learn protein sequences. It's harder to learn protein structures. Suppose you discover a new type of protein structure, call it a *helicoil h*. It's rare; you've searched 5000 proteins and found only 20 helicoils, so $p(h) = 0.004$. If you could discover some special amino acid *sequence feature*, call it sf, that predicts the h structure, you could search other genomes to find other helicoil proteins in nature. It's easier to turn this around. Rather than looking through 5000 sequences for patterns, you want to look at the 20 helicoil proteins for patterns. How do you compute $p(\text{sf} \mid h)$? You take the 20 given helicoils and find the fraction of them that have your sequence feature. If your sequence feature (say alternating glycine and lysine amino acids) appears in 19 out of the 20 helicoils, you have $p(\text{sf} \mid h) = 0.95$. You also need $p(\text{sf} \mid \bar{h})$, the fraction of non-helicoil proteins (let's call those $\bar{h}$) that have your sequence feature. Suppose you find $p(\text{sf} \mid \bar{h}) = 0.001$. Combining Equations (1.11) and (1.15) gives Bayes' rule for the probability you want:

$$p(h \mid \text{sf}) = \frac{p(\text{sf} \mid h)p(h)}{p(\text{sf})} = \frac{p(\text{sf} \mid h)p(h)}{p(\text{sf} \mid h)p(h) + p(\text{sf} \mid \bar{h})p(\bar{h})}$$
$$= \frac{(0.95)(0.004)}{(0.95)(0.004) + (0.001)(0.996)} = 0.79. \qquad (1.16)$$

In short, if a protein has the sf sequence, it will have the h structure about 80% of the time.

example of bayesian classifier (Bulashevska2008) Why not to try?

Journal of Theoretical Biology 254 (2008) 799–803

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered

Alla Bulashevska [a,*], Roland Eils [a,b]

[a] Department of Theoretical Bioinformatics, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[b] Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology (IPMB), University of Heidelberg, Germany

ARTICLE INFO

Article history:
Received 19 November 2007
Received in revised form
19 May 2008
Accepted 19 May 2008
Available online 14 June 2008

Keywords:
Unfolded proteins
Disorder prediction
Model-based classification
Multinomial model

ABSTRACT

Intrinsically disordered proteins (IDPs) lack a well-defined three-dimensional structure under physiological conditions. Intrinsic disorder is a common phenomenon, particularly in multicellular eukaryotes, and is responsible for important protein functions including regulation and signaling. Many disease-related proteins are likely to be intrinsically disordered or to have disordered regions. In this paper, a new predictor model based on the Bayesian classification methodology is introduced to predict for a given protein or protein region if it is intrinsically disordered or ordered using only its primary sequence. The method allows to incorporate length-dependent amino acid compositional differences of disordered regions by including separate statistical representations for short, middle and long disordered regions. The predictor was trained on the constructed data set of protein regions with known structural properties. In a Jack-knife test, the predictor achieved the sensitivity of 89.2% for disordered and 81.4% for ordered regions. Our method outperformed several reported predictors when evaluated on the previously published data set of Prilusky et al. [2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21 (16), 3435–3438]. Further strength of our approach is the ease of implementation.

A B S T R A C T

Intrinsically disordered proteins (IDPs) lack a well-defined three-dimensional structure under physiological conditions. Intrinsic disorder is a common phenomenon, particularly in multicellular eukaryotes, and is responsible for important protein functions including regulation and signaling. Many disease-related proteins are likely to be intrinsically disordered or to have disordered regions. In this paper, a new predictor model based on the Bayesian classification methodology is introduced to predict for a given protein or protein region if it is intrinsically disordered or ordered using only its primary sequence. The method allows to incorporate length-dependent amino acid compositional differences of disordered regions by including separate statistical representations for short, middle and long disordered regions. The predictor was trained on the constructed data set of protein regions with known structural properties. In a Jack-knife test, the predictor achieved the sensitivity of 89.2% for disordered and 81.4% for ordered regions. Our method outperformed several reported predictors when evaluated on the previously published data set of Prilusky et al. [2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21 (16), 3435–3438]. Further strength of our approach is the ease of implementation.

https://www.disprot.org/about
https://protein.bio.unipd.it/projects
https://proteopedia.org/wiki/index.php/Main_Page

The focus of the paper

In this paper, we introduce a new prediction method, which exploits the Bayesian classification procedure to predict disordered property for a given protein or protein region from its primary sequence. Bayesian Markov chain model-based classification has already found its application in proteomics for the prediction of protein subcellular locations (Bulashevska and Eils, 2006). This approach represents each class with a single probabilistic summary. Since the AAC of disordered regions is distinct from that of ordered, we propose to use multinomial models for the description of class-conditional densities. The intuition behind this approach is that each protein sequence belonging to a certain class can be considered as a realization of an independent random process that emits symbols from an alphabet of 20 amino acids.

2.2. Multinomial models

Multinomial models assume a *bag-of-amino acid* sequence representation, which considers the appearance of each amino acid as an independent event. The order in which amino acids occur in a given amino acid sequence is ignored; the only information retained is a vector of counts $\mathbf{n} = (n_1, \ldots, n_{20})$, where n_i is the number of occurrences of amino acid i in the sequence.

We assume that the probability of a sequence s to come from a certain class c is given by a multinomial probability function governed by its vector of parameters $\theta_c = (\theta_{c1}, \ldots, \theta_{c20}) \in [0, 1]^{20}$:

$$p(s|\theta_c) = \frac{n!}{\prod_{i=1}^{20} n_i!} \prod_{i=1}^{20} \theta_{ci}^{n_i}, \tag{1}$$

where $n = \sum_i n_i$ denotes the length of the sequence. The parameter θ_{ci} denotes the cth class-conditional probability of amino acid i to occur in a sequence. The parameters of the model corresponding to class c are estimated from the training regions belonging to the class c. Thus, the parameter θ_{ci} is calculated as

$$\theta_{ci} = \frac{n_{ci}}{\sum_{i=1}^{20} n_{ci}}, \tag{2}$$

where n_{ci} is the number of occurrences of amino acid i in the sequences of class c. This way of estimating parameters of the

2.3. Bayesian multinomial classifier

Bayesian classification is a widely applied method in the machine learning and statistical community, which is based on Bayes' theorem (Bayes rule). According to Bayes' rule, the class for an unlabeled sequence s can be inferred using the posterior probability:

$$p(c|s) = \frac{p(c)p(s|c)}{p(s)} = \frac{p(c)p(s|c)}{\sum_c p(c)p(s|c)}. \quad (3)$$

Note!

We assume class prior probabilities $p(c)$ to be equally distributed. We further assume that the sequences of each class are generated from multinomial models. Thus, given the parameters $\{\theta_c\}$ of the models for each class, the term $p(s|c)$ denoting the prior probability of a sequence s to belong to the class c can be computed using the formula (1) for $p(s|\theta_c)$ from previous subsection.

Since we model short, middle and long disordered regions separately, the estimation of the class-conditional densities involves four subproblems (for short, middle, long disordered and ordered classes), in which each of the class-conditional density is estimated based on the data belonging to the corresponding class only.

Bayesian classifier is a *probabilistic* classifier, which yields for each query instance the posterior probability for each class, a numeric value that represents the degree to which an instance is a member of a class. To produce a discrete output, the following decision rule is usually applied: the class should be the one which maximizes the posterior probability.

NOTE!

To classify an input sequence as disordered or ordered, we sum the posterior probabilities for short, middle and long disordered subtypes into a single value describing the posterior probability of a sequence to be disordered and then use the standard decision rule to come up with a discrete output, i.e. predict one of the two classes (disordered/ordered) showing the biggest posterior probability.

2.4. Performance evaluation

The prediction performance of our predictor was validated with *Jack-knife test* (or *leave-one-out cross-validation*) (Mardia et al., 1979). By Jack-knife test the learning step is performed with all training instances except the one for which the class is to be predicted.

The prediction quality was evaluated using the standard measures of *sensitivity* (SN) and *specificity* (SP), where the sensitivity, or *true positive rate*, is the percentage of disordered sequences correctly predicted, and the SP, or *true negative rate*, is the percentage of ordered sequences correctly predicted. We calculate the *overall accuracy* (ACC) as the average of SN and SP, which is more suitable than the percentage of all correctly predicted sequences for data sets with imbalanced class distributions. We also show receiver operating characteristic (ROC) curve and report area under the ROC curve (AUC) calculated using the R package ROCR (Sing et al., 2005).

Jackknife

One of the earliest techniques to obtain reliable statistical estimators is the jackknife technique. It requires less computational power than more recent techniques.

Suppose we have a sample $x=(x_1,x_2,...,x_n)$ and an estimator $\hat{\theta}=s(x)$. The jackknife focuses on the samples that *leave out one observation at a time*:

$$x_{(i)}=(x_1,x_2,...,x_{i-1},x_{i+1},...,x_n)$$

for $i=1,2,...,n$, called *jackknife samples*. The *i*th jackknife sample consists of the data set with the *i*th observation removed. Let $\hat{\theta}_{(i)}=s(x_{(i)})$ be the *i*th jackknife replication of $\hat{\theta}$. The jackknife estimate of standard error defined by

$$\widehat{SE}_{jack}=[\frac{n-1}{n}\sum(\hat{\theta}_{(i)}-\hat{\theta}_{(.)})^2]^{1/2} \qquad (3)$$

where $\hat{\theta}_{(.)}=\sum_{i=1}^{n}\hat{\theta}_{(i)}/n$.

The jackknife only works well for linear statistics (e.g., mean). It fails to give accurate estimation for non-smooth (e.g., median) and nonlinear (e.g., correlation coefficient) cases. Thus improvements to this technique were developed.

https://www.datasciencecentral.com/profiles/blogs/resampling-methods-comparison

True Positive:
Interpretation: You predicted positive and it's true.
You predicted that a woman is pregnant and she actually is.
True Negative:
Interpretation: You predicted negative and it's true.
You predicted that a man is not pregnant and he actually is not.
False Positive: (Type 1 Error)
Interpretation: You predicted positive and it's false.
You predicted that a man is pregnant but he actually is not.
False Negative: (Type 2 Error)
Interpretation: You predicted negative and it's false.
You predicted that a woman is not pregnant but she actually is.

Indicators to evaluate methods

$$\text{Sensitivity (or recall)} : S_n = \frac{TP}{TP + FN} = \frac{TP}{N_d} \tag{1}$$

is the number of correctly identified disordered proteins normalized to the total number of disordered proteins in the sample

$$\text{Specificity} : S_p = \frac{TN}{TN + FP} = \frac{TN}{N_o} \tag{2}$$

is the ratio between the number correctly identified ordered proteins and the total number of ordered proteins in the sample;

$$\text{Rate of false positives} : f_p = \frac{FP}{TN + FP} = 1 - S_p \tag{3}$$

is the ratio between the number of ordered proteins predicted as disordered and the total number of ordered proteins in the sample;

$$\text{Accuracy} : ACC = \frac{S_n + S_p}{2} \tag{4}$$

that is the average between sensitivity and specificity. It measures the overall performance of the predictor. Then,

$$\text{Precision (or selectivity)} : \text{Pr} = \frac{TP}{TP + FP} = \frac{TP}{n_d} \tag{5}$$

Model comparison by Bayes factors

A model M_i, in the Bayesian sense, is a pair consisting of a conditional likelihood function $P(D \mid \theta, M_i)$ for observable data D together with a prior $P(\theta \mid M_i)$ over parameter vector θ. Ideally, we might like to assess the absolute probability of model M_i after seeing data D. We can express this quantity using Bayes rule:

$$P(M_i \mid D) = \frac{P(M_i)\, P(D \mid M_i)}{\int P(M_j)\, P(D \mid M_j)\, \mathrm{d}M_j} \,.$$

Quantifying over the whole class of possible models is dauntingly complex. This likely remains true even for a fixed data set and within a confined sub-genre of models (e.g., all regression models with combinations of a finite set of explanatory factors).

A good solution is to be more modest and to compare just two models to each other. The question to ask is then: How much does the observation of data D impact our beliefs about the relative model probabilities? We can express this using Bayes rule as the ratio of our posterior beliefs about models, which eliminates the need to have the normalizing constant for the previous equation, like so:

$$\underbrace{\frac{P(M_1 \mid D)}{P(M_2 \mid D)}}_{\text{posterior odds}} = \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}} \underbrace{\frac{P(D \mid M_1)}{P(D \mid M_2)}}_{\text{Bayes factor}} \,.$$

The fraction on the left-hand side is the posterior odds ratio: our relative beliefs about models M_1 and M_2 *after* seeing data D. On the right-hand side we have a product of two intuitively interpretable quantities. First, there is the prior odds: our relative beliefs about models M_1 and M_2 *before* seeing data D. Second, there is the so-called **Bayes factor**. This way of introducing Bayes factors invites to think of them as **the factor by which our prior odds change in the light of the data**.

synthesis try computer codes at: https://michael-franke.github.io/statistics,/modeling/2017/07/07/BF_computation.html

BIOINFORMATICS

THE MACHINE LEARNING APPROACH

SECOND EDITION

PIERRE BALDI AND SØREN BRUNAK

Appendix B

Information Theory, Entropy, and Relative Entropy

Here we briefly review the most basic concepts of information theory used in this book and in many other machine learning applications. For more in-depth treatments, the reader should consult [483], [71], [137], and [577]. The three most basic concepts and measures of information are the entropy, the mutual information, and the relative entropy. These concepts are essential for the study of how information is transformed through a variety of operations such as information coding, transmission, and compression. The relative entropy is the most general concept, from which the other two can be derived. As in most presentations of information theory, we begin here with the slightly simpler concept of entropy.

B.1 Entropy

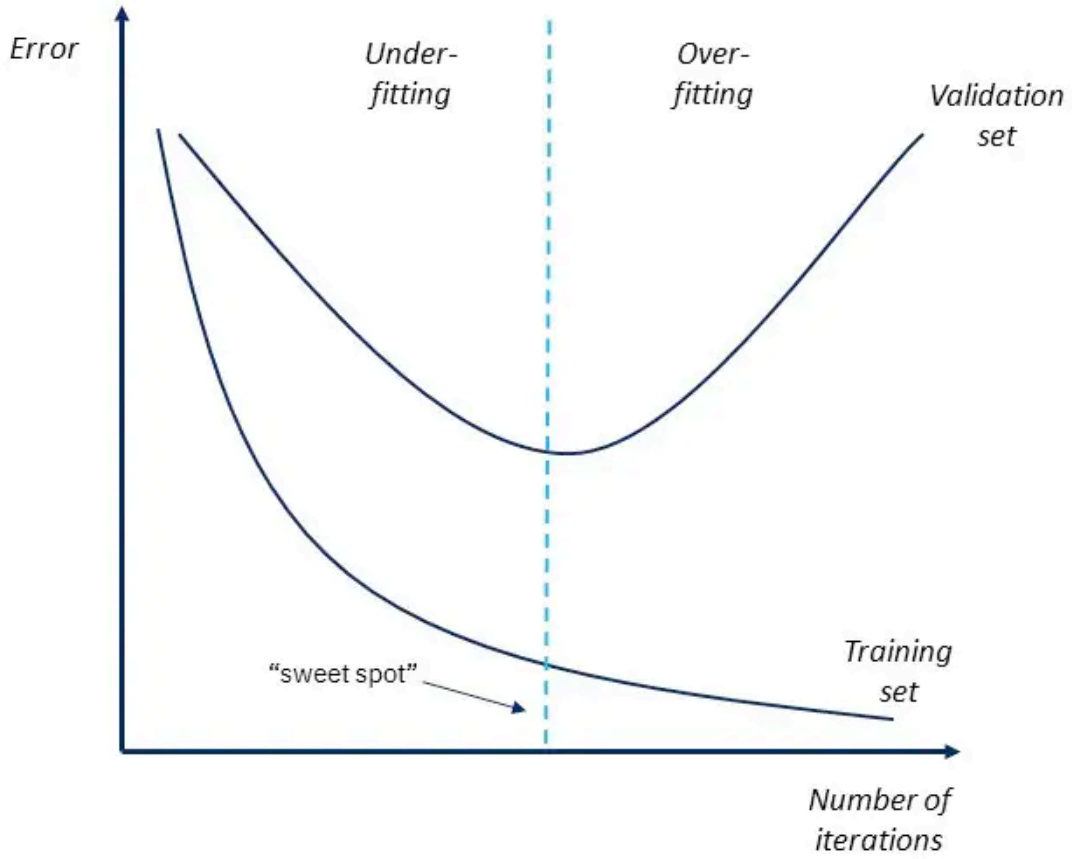
The entropy $\mathcal{H}(P)$ of a probability distribution $P = (p_1, \ldots, p_n)$ is defined by

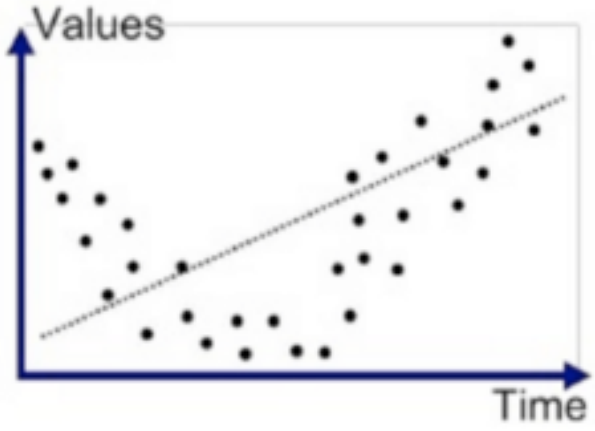
$$\mathcal{H}(P) = \mathbf{E}(-\log P) = -\sum_{i=1}^{n} p_i \log p_i. \tag{B.1}$$

The units used to measure entropy depend on the base used for the logarithms. When the base is 2, the entropy is measured in bits. The entropy measures the prior uncertainty in the outcome of a random experiment described by P, or the information gained when the outcome is observed. It is also the minimum average number of bits (when the logarithms are taken base 2) needed to transmit the outcome in the absence of noise.

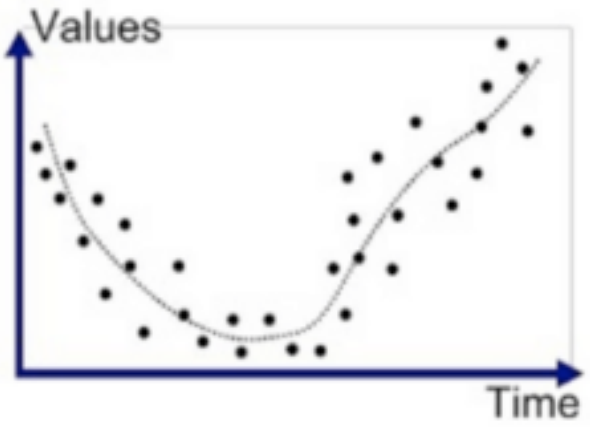
Homework complete the study of Appendix E in Baldi and Brunak's textbook and prepare an exposition

Underfitting/Overfitting

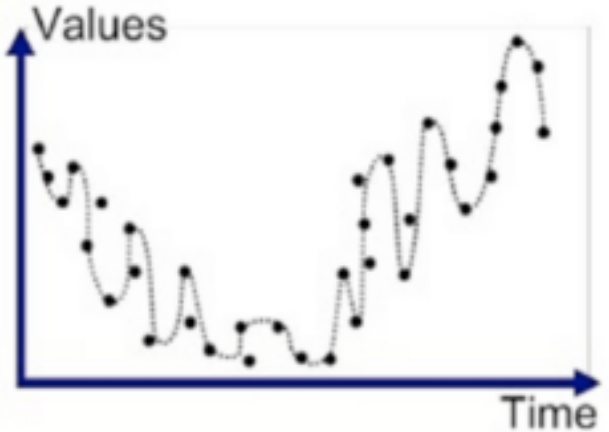




Underfitted



Good Fit/Robust



Overfitted