

# Extended Abstract: Maximum Expected Utility Modeling of the Conditional Probability of Adenocarcinoma

Craig Friedman  
Courant Institute, NYU

251 Mercer Street  
New York, NY  
(212)724-2109

craig.friedman@cims.nyu.edu

Wenbo Cao  
CUNY Graduate Center  
365 Fifth Avenue  
New York, NY 10016

(212)678-4977

wcao@gc.cuny.edu

## ABSTRACT

There have been a number of quantitative methods applied to the classification and clustering of microarray data (see, for example, [6]). In this article, we describe a statistical learning theory-based method to model the probability of adenocarcinoma, conditioned on gene expression microarray data. We find the conditional probability distribution by choosing a model on an efficient frontier, which we define in terms of consistency with the training data and consistency with a prior distribution. We measure the former by means of the large-sample distribution of a vector of sample-averaged features, and the latter by means of the relative entropy. This formulation leads to an optimization problem for which the dual problem's objective function is, under certain conditions, asymptotically, the expected utility for an investor who uses the model to bet on whether particular samples are adenocarcinoma or not. We measure the performance of the models that we create in economic terms. We consider a popular benchmark model, the linear logit (also known as logistic regression), which we describe. In our numerical experiments, our model has clearly outperformed the linear logit model. We used the 58 probe set identifiers common to both the Harvard and Michigan data sets. When we trained on either data set, we were able to perfectly classify set adenocarcinoma on the other set.

## General Terms

Algorithms, Measurement, Performance, Human Factors

## Keywords

Microarray, ontology, adenocarcinoma, conditional probability

## 1. THEORETICAL FRAMEWORK

### 1.1 Conditional Probabilities

Let  $x$  denote our vector of explanatory variables and the random variable  $Y \in \{0,1\}$  indicate adenocarcinoma ( $Y=1$ ) or non-adenocarcinoma ( $Y=0$ ) over some fixed time interval from the observation of  $x$ . We seek the conditional probability measure  $p(y|x) = \text{Prob}(Y=1|x)$ .

### 1.2 Maximum Expected Utility Principle and Dual Problem

We follow the modeling approach from [3], who seek a probability measure that maximizes the out-of-sample expected utility of an investor who chooses his investment strategy so as to maximize his expected utility under the model he believes. It is assumed here that, for each value of  $x$ , the investor distributes his wealth over bets on both possible outcomes, adenocarcinoma and or non-adenocarcinoma, where each bet has an associated payoff (or odds ratio or price) in a hypothetical market.

The approach from [3] asymptotically achieves this goal by selecting the model with the best out-of-sample expected utility among the models on an efficient frontier. This efficient frontier is defined as the set of all (Pareto optimal) measures for which there is no measure that is simultaneously more consistent with the data. We measure consistency with the data via the difference between the theoretical and sample expectations of a set of features. We may view a feature as a mapping from the pairs  $(y,x)$  to the real numbers (see, for example, [5]), and with our prior (the model that we believe  $\{it\}$  before we observe data) measure. The measures on the efficient frontier are parameterized by the single parameter  $\alpha$ .

Here, we use this approach with a logarithmic utility function. It has been shown by [2] that the assumption of a logarithmic utility can lead to approximately expected-utility-optimal models, even for investors with non-logarithmic utilities, under plausible assumptions. This means that, for each value  $\alpha$ , we find the Pareto optimal measure by minimizing, over measures  $p$ , the discrepancy (Kullback-Leibler relative entropy) between  $p$  (the measure that we seek) and the prior  $p^0$ :

$$D(p\|p^0) = \sum_x \tilde{p}(x) \sum_{y=0,1} p(y|x) \log \frac{p(y|x)}{p^0(y|x)}, \quad (1)$$

$$\text{subject to } Nc^T \Sigma^{-1} c \leq \alpha, \quad (2)$$

$$\text{with } c = E_p[f] - E_{\tilde{p}}[f], \quad (3)$$

$$E_p[f] = \sum_x \tilde{p}(x) \sum_{y=0,1} p(y|x) f(y, x) \quad (4)$$

$$\text{and } E_{\tilde{p}}[f] = \sum_x \tilde{p}(x) \sum_{y=0,1} \tilde{p}(y|x) f(y, x). \quad (5)$$

Here,  $f(y,x)=(f_1(y,x),\dots,f_i(y,x))^T$  is the vector of features,  $p$  denotes the empirical distribution, and  $\Sigma$  is the empirical covariance matrix of the features. The notion of a feature is used heavily in the statistical learning community but is not as widely used in finance at present. The definition for a feature is rather abstract: it is a mapping from the pairs  $(x,y)$  to the real numbers. As far as the role played by the features goes, the introduction of more and more features allows for more and more feature constraints, which allows for more and more consistency of the model with the data. The solutions to above optimization problems form a family of measures which is parameterized by  $\alpha$ . After computing a number of these Pareto optimal measures, we will pick the one that has the best out-of-sample performance as measured by the expected utility (see Section 3).

In practice, we take

$$\tilde{p}(x) = \frac{1}{N} \quad (6)$$

in the above sums, where  $N$  is the number of observations in the training data set, and

$$\tilde{p}(y|x) \in \{0, 1\}. \quad (7)$$

We can measure the quality of a model  $p$  by the relative outperformance of the model over the benchmark (prior) model,  $p^0$ : the gain in expected (under the unknown true measure) utility experienced by an investor who invests optimally according to the model relative to an investor who invests optimally according to the benchmark model  $p^0$  (see [2]). It can be shown (see [3]) that every Pareto optimal measure is robust in the sense that it maximizes, over all measures, the worst-case (over potential true measures equally consistent with the data) relative outperformance of the model over the benchmark model. Alternatively, we may interpret the above approach as a minimum relative entropy approach, similar to the one of [4]. To this end we adopt the point of view that our probability measure should contain as little information beyond the prior information as possible while being consistent with the feature expectation constraints; we measure information by means of the relative entropy (1). Allowing  $\alpha > 0$  in the constraint (2) regularizes the probability density, i.e. it allows for an imperfect match of the model and empirical features averages. This mitigates overfitting.

It is well known (see, for example[5]) that the duals of some relative entropy minimization problems are maximum likelihood problems for an exponential distribution. This is also true for the problem discussed above, as was shown by [3]. To be specific, the dual of above optimization problem is the following:

$$\text{Find } \beta^* = \arg \max_{\beta} h(\beta) \quad (8)$$

$$\text{with } h(\beta) = \frac{1}{N} \sum_{k=1}^N \log p^{(\beta)}(y_k|x_k) - \sqrt{\frac{\alpha}{N}} \beta^T \Sigma \beta \quad (9)$$

$$\text{with } p^{(\beta)}(y|x) = \frac{1}{Z_x(\beta)} e^{\beta^T f(y,x)} p^0(y|x) \quad (10)$$

$$\text{and } Z_x(\beta) = \sum_{y=0,1} p^0(y|x) e^{\beta^T f(y,x)}, \quad (11)$$

where the  $(x_k, y_k)$  are the observed  $(x,y)$ -pairs,  $N$  is the number of observations, and  $\beta=(\beta_1,\dots,\beta_j)^T$  is a parameter vector. Our optimal (in the sense of the dual and of our original optimization problem) probability measure is then:

$$p(y|x) = p^{(\beta^*)}(y|x) . \quad (12)$$

The problem (8)-(11) (see [3]) amounts to a regularized maximum likelihood estimation, or an expected utility maximization, of the parameter vector  $\beta$ , where  $p^{(\beta^*)}(y|x)$  has an exponential factor. We will use this formulation to numerically find this optimal measure. For a practical implementation, we can drop the square root in the second term of (9); the resulting family, indexed by  $\alpha$ , of solutions is the same (see [3]). The objective function of the maximization problem (8)-(11) is strictly concave. For this reason, the problem is amenable to a robust numerical solution.

For the data we worked with, the results depend little on the prior measure; the results described below were obtained for the following convenient choice:

$$p^0(Y = 1|x) = \frac{\sum_{\{k:y_k=1\}} 1}{N}, \quad (13)$$

i.e., the prior distribution is flat with conditional probability of adenocarcinoma equal to the unconditional probability of adenocarcinoma.

### 1.3 Features

We use three types of features:

(i) Linear features

$$f(y, x) = (y - \frac{1}{2})(x)_j \quad (14)$$

where  $(x)_j$  denotes the  $j^{\text{th}}$  coordinate of  $x$ , with the convention that  $(x)_0 = 1$ .

(ii) Quadratic features

$$f(y, x) = (y - \frac{1}{2})(x)_i(x)_j, \text{ and} \quad (15)$$

(iii) Gaussian kernel features

$$f(y, x) = (y - \frac{1}{2})exp(-\|x - x_k\|^2) \quad (16)$$

where  $x_k$  is an observed value of  $x$ .

## 2. THE BENCHMARK MODEL: LINEAR LOGISTIC REGRESSION

The linear logistic regression is given by

$$p(1|x) = \frac{1}{1 + e^{-\sum_i \beta_i x_i}}, \quad (17)$$

where the parameters,  $\beta_i$ , are chosen to maximize the likelihood function.

We note that linear logistic regression is a special case of our approach when the prior is flat,  $a=0$  and

$$f_j(y, x) = (y - \frac{1}{2})(x)_j. \quad (18)$$

Thus, our approach can be viewed as a generalization which is better able to handle nonlinearities and overfitting.

It is easy to show that, for the linear logit model, the level sets of the conditional probability of adenocarcinoma surface must satisfy a rather strict geometrical condition: they must be linear. This imposes a severe restriction on the model, which may not be sufficiently flexible to conform to the story told by the data.

## 3. MODEL PERFORMANCE MEASUREMENT

To measure the performance of our model,  $p$ , we use the scaled log-likelihood difference between our model and the (non-informative) prior measure, as estimated on an out-of-sample dataset. This performance measure is computed as:

$$\Delta_{log}(p) = \frac{1}{N} \sum_{k=1}^N \log p^{(\beta^*)}(y_k|x_k), \quad (19)$$

where the  $(x_k, y_k)$  are the  $(x, Y)$ -pairs and  $N$  the number of observations of the test sample. It was shown by [2] that this

quantity can be interpreted as the gain in expected logarithmic utility experienced by an investor who uses the model  $p$  to design a utility-optimal investment strategy, where the gain is measured with respect to an investor who has no information and therefore invests according to the prior measure,  $p^0$ . The above measure is equal to the pickup in the expected wealth growth rate for the investor who invests according to the model  $p$  rather than  $p^0$  (see, for example, [1] or [2]). To be specific, we define the wealth growth rate, for an investment strategy based on a given model  $q$  by means of

$$g(q) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{W_n(q)}{W_0} \quad (20)$$

where  $W_n(q)$  is the investor's wealth after  $n$  bets. The difference in the growth rates between a model- $p$ -based and a prior-measure-based investment strategy is then

$$g(p) - g(p^0) = \Delta_{log}(p), \quad (21)$$

(see [2]). Note that we compute only the *relative* wealth growth rate, i.e., the difference in the wealth growth rates for the investment strategies based on two different models.

To compute our performance measure, we train on one of the data sets (Harvard or Michigan) and calculate the performance measure on the other.

For each model, we also display the following popular, rank-based performance measures: Receiver Operator Characteristic (ROC) curve. For definitions, references and a discussion of the relative merits of these rank-based model performance measures, [2].

## 4. RESULTS

Training on Harvard data, Testing on Michigan data, using all 58 variables:

MEU method: roc = 1, delta = 0.0579

Linear Logit: roc = 1, delta = 0.0579

Using Michigan as Training data, Using Harvard as Testing data, all 58 variables:

MEU method: roc = 1, delta = 0.3433

Linear Logit: roc = 0.8667, delta = -1.6886

We note that the MEU model produced perfect classification on the out of sample data sets.

## 5. REFERENCES

- [1] Cover, T., and Thomas, J. Elements of Information Theory, John Wiley & Sons, Inc., 1991.
- [2] Friedman, C. and Sandow, S. Model Performance Measures for Expected Utility Maximizing Investors. International Journal of Theoretical and Applied Finance, June, 2003.
- [3] Friedman, C., and Sandow, S. Learning Probabilistic Models: An Expected Utility Approach. Journal of Machine Learning Research. July, 2003.
- [4] Jaynes, E. Information Theory and Statistical Mechanics. Physical Review, 106, 1957, p.620.
- [5] Lebanon, G., and Lafferty, J. Boosting and Maximum Likelihood for Exponential Models. Advances in Neural Information Processing Systems (NIPS), 14, 2001.
- [6] Lin., Simon and Johnson, K. Methods of Microarray Data Analysis II, Kluwer, 2002.