Stat 538 Project:
Implementation of Maximum Entropy Discrimination
with a Linear Discriminant Function

Jonathan Gruhl
March 16, 2008

1 Introduction

For this project, our goal was to implement Maximum Entropy Discrimination (MED) with a linear discrimination function for binary classification purposes. This technique was presented by Jaakola, Meila, and Jebara in “Maximum Entropy Discrimination” (1999). Their paper provides a generalized framework for discrimination that relies on the maximum entropy principle. One of the interesting aspects of MED is its consideration of distributions over parameters in the discriminative model. Thus, instead of finding a single optimal setting of parameters, MED estimates a distribution over the parameters and this distribution is chosen to satisfy constraints in a manner that is non-prejudiced or does not imply that possession of information not present. We apply this technique with a linear discriminant function first to a common benchmarking dataset for discrimination techniques and second to a classification problem of interest to the author.

The benchmark dataset is the common SONAR, Mines vs. Rocks dataset. This is one of the datasets used in “Benchmarking Support Vector Machines” by Meyer, Leisch & Hornik (2002). This dataset was originally used in “Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets” by Gorman and Sejnowski. Their task was to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock. This dataset has since been used by many others and we intend to compare our average classification rates to other linear classifiers to ensure that the MED classifier is performing adequately. Although we expect some classifiers to outperform our MED linear classifier, we expect the (correctly implemented - hopefully) MED linear classifier to demonstrate average classification rates consistent or superior to those of a linear Support Vector
Machine (SVM). This dataset consists of 208 data points with 60 inputs and the binary outcome of “R” for rock and “M” for mine.

The second task to which we apply MED is the classification of win and loss outcomes for a professional basketball team. Beyond arcane personal interest, what value does this information have? Win/loss contracts are one of the traded individual game contracts on sports trading exchanges such as TradeSports that have become increasingly popular for professional and amateur gamblers. The point spread for the game in question is the natural predictor on which to base the classification of a win/loss outcome. A point spread is the difference between the expected score of one team and the expected score of the opposing team. Point spreads are the mechanism by which sports books ensure that they even amounts of money committed to opposing sides of a bet. Like market prices for a stock, a point spread for a game is often assumed to incorporate all public information about the outcome of the game. The direction of the point spread could reasonably considered as the best predictor of the win/loss outcome for a basketball game with the magnitude of the point spread serving as an indicator of probability of win/loss.

Still, the function of the point spread is not to directly predict the win/loss outcome of a game. In our example here, we consider some additional items of information to see if we can improve upon the classification accuracy of the point spread. The additional predictors are days of rest for a team, days of rest for the opposing team, and the over/under for the game. As games for professional basketball teams are not evenly spaced, days of rest is usually imbalanced for the two teams and the team with more rest could be expected to retain an advantage. The over/under of a game is the expected total of the two team’s scores. Thus, it can be expected to reflect “market” information about the anticipated nature of the game. A high over/under would suggest a fast-paced and high-scoring game while a low over /under would suggest the opposite. Presumably this information is reflected in the point spread but to what extent is unknown. To examine whether this additional information can supplement the point spread in the accuracy of win/loss classification, we use the point spreads, over/unders and days of rest for 400 Washington Wizards games spanning the 2003-2008 seasons.

In the next section, we summarize the MED technique and present expressions for its application when using a linear discriminant function. Following that, we provide results of the two above experiments.

2 Maximum Entropy Discrimination

In this section we describe the Maximum Entropy Discrimination technique. We follow the expositions of Jaakkola, Meila, and Jebara [JMJ] and Jebara [J] closely, briefly introducing their main results related to linear discrimination
while also presenting expressions for the dual and predicted labels specific to a linear discriminant function. (Although some details of the derivations of the expressions are present in J, they have been re-derived for the author’s own educational benefit and presented here in full).

If we have a classification situation with two outcome classes, \( y \in \{-1, 1\} \) for inputs \( X \), then the assumption of class-conditional probabilities over the inputs could be encapsulated in the discriminant function \( L(X|\Theta) = \log \frac{P(X|\theta_1, y=1)}{P(X|\theta_{-1}, y=-1)} + b \) where \( b \) is a bias term and \( \Theta = \{\theta_1, \theta_{-1}, b\} \). Instead of determining single values for each of these parameters in \( \Theta \), JMJ consider the problem of finding a distribution \( P(\Theta) \) over the parameters. They then replace the discriminant function \( L(X|\Theta) \) with \( \int P(\Theta) L(X|\Theta) d\Theta \).

How might we choose this distribution \( P(\Theta) \)? JMJ use a maximum entropy framework where they maximize the entropy \( H(P) \) subject to classification constraints \( \int P(\Theta)[y_t L(X|\Theta)] d\Theta > \gamma \) where \( \gamma \) is an intended classification margin. As J notes, the entropy can be cast as a Kullback-Leibler divergence from a prior uniform distribution (or another distribution to reflect any prior knowledge we have). This leads us to the following definition of the Maximum Entropy Discrimination solution.

**Definition 1** The MED solution is the \( P(\Theta, \gamma) \) over the parameters \( \theta \) and the margin variables \( \gamma = [\gamma_1, \ldots, \gamma_T] \) that minimizes \( KL(P_\theta \parallel P_\theta^0) + \sum_t KL(P_{\gamma_t} \parallel P_{\gamma_t}^0) \) subject to \( \int P(\Theta, \gamma)[y_t L(X|\Theta) - \gamma_t] d\Theta d\gamma \geq 0 \forall t \). Here \( P_\theta^0 \) is the prior distribution over the parameters and \( P_{\gamma_t}^0 \) is the prior over margin variables. The resulting decision rule is given by \( \hat{y} = \text{sign}(\int P(\Theta)L(X|\Theta)d\Theta) \).

As discussed in class and in J, the KL divergence terms are convex functions of \( P(\Theta) \) and \( P(\gamma_t) \). Additionally, JMJ note that if it is assumed that there is a non-zero probability for all \( \gamma_t \) taking negative values, the admissible set of distributions \( P(\Theta, \gamma) \) meeting the classification constraints is never empty. As a result, even if the training examples cannot be separated by any discriminant function in our linear class, a valid and unique solution is still available. JMJ then arrive at the following solution to the non-separable MED classification problem in Definition 1. J notes that this is the dual to the constrained optimization problem listed in Definition 1.

**Theorem 1** The solution to the MED problem for estimating a distribution over parameters and margins (as well as further augmentations) has the following general form:

\[
P(\Theta, \gamma) = \frac{1}{Z(\lambda)} P_0(\Theta, \gamma) e^{\sum_t \lambda_t [y_t L(X|\Theta) - \gamma_t]}
\]

where \( Z(\lambda) \) is the normalization constant (partition function) and \( \lambda = \{\lambda_1, \ldots, \lambda_T\} \) defines a set of non-negative Lagrange multipliers, one for each classification.
constraint. \( \lambda \) are set by finding the unique maximum of the jointly concave objective function:

\[
J(\lambda) = -\log Z(\lambda)
\]

Thus finding the solution depends on being able to evaluate

\[
Z(\lambda) = \int P_0(\Theta, \gamma) e^{\sum \lambda_t[y_t L(X|\Theta) - \gamma_t]} d\Theta d\gamma
\]

Here we concern ourselves with the linear discriminant boundary \( L(X|\Theta) = \theta^T X + b \). For our prior for \( \gamma \), we assume \( P_0(\gamma) = \prod_t P_0(\gamma_t) \) and as in JMJ

\[
P_0(\gamma_t) = ce^{-c(1-\gamma_t)} \text{ for } \gamma_t \leq 1.
\]

For the prior for the bias term \( b \), we employed a \( N(0, \sigma_b^2) \) distribution. This prior will favor bias values close to zero so we choose a fairly large value of \( \sigma_b^2 \) so as to still permit a reasonably wide range of values for \( b \). If we further assume \( P_0(\theta) \) is distributed \( N(0, I) \) and \( \theta \) independent from \( b \), then

\[
Z(\theta) = \int P_0(\Theta, \gamma) \exp \left( \sum \lambda_t[y_t(\theta^T x_t + b) - \gamma_t] \right) d\Theta d\gamma
\]

\[
= \int P_0(\theta) \exp \left( \sum \lambda_t y_t \theta^T x_t \right) P_0(b) \exp \left( \sum \lambda_t b \right) P_0(\gamma) \exp \left( -\sum \lambda_t \gamma_t \right) d\theta db d\gamma
\]

If we consider each element individually

\[
\int P_0(\gamma_t) \exp(-\lambda_t \gamma_t) d\gamma_t = \int_{-\infty}^{1} c \exp(-c + c\gamma_t) \exp(-\lambda_t \gamma_t) d\gamma_t
\]

\[
= \int_{-\infty}^{1} c \exp(-c + c\gamma_t - \lambda_t \gamma_t) d\gamma_t
\]

\[
= \frac{c}{c - \lambda_t} \exp(-\lambda_t) = \frac{1}{1 - \frac{\lambda_t}{c}} \exp(-\lambda_t)
\]

\[
\int P_0(b) \exp \left( \sum \lambda_t y_t b \right) db = \int_{-\infty}^{\infty} \left(2\pi \sigma_b^2\right)^{-1/2} \exp \left( -\frac{b^2}{2\sigma_b^2} \right) \exp \left( b \sum \lambda_t y_t \right) db
\]

\[
= \int_{-\infty}^{\infty} \left(2\pi \sigma_b^2\right)^{-1/2} \exp \left( -\frac{1}{2} \left( \frac{b^2}{\sigma_b^2} - 2b \sum \lambda_t y_t \right) \right) db
\]

\[
= \exp \left( \frac{1}{2} \sigma_b^2 \left( \sum \lambda_t y_t \right)^2 \right) \int_{-\infty}^{\infty} \left(2\pi \sigma_b^2\right)^{-1/2} \exp \left( -\frac{1}{2\sigma_b^2} \left( b - \sigma_b^2 \sum \lambda_t y_t \right)^2 \right) db
\]

\[
= \exp \left( \frac{1}{2} \sigma_b^2 \left( \sum \lambda_t y_t \right)^2 \right)
\]
\[
\int P_0(\theta) \exp \left( \sum_t \lambda_t y_t \theta^T x_t \right) d\theta = \int_{-\infty}^{\infty} (2\pi)^{-n/2} \exp \left( -\frac{1}{2} \theta^T \theta \right) \exp \left( \sum_t \lambda_t y_t \theta^T x_t \right) d\theta \\
= \int_{-\infty}^{\infty} (2\pi)^{-n/2} \exp \left( -\frac{1}{2} \left( \theta^T \theta - 2\theta^T \sum_t \lambda_t y_t x_t \right) \right) d\theta \\
= \exp \left( \frac{1}{2} \left( \sum_t \lambda_t y_t x_t \right)^T \left( \sum_t \lambda_t y_t x_t \right) \right) \\
\int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} \left( \theta - \sum_t \lambda_t y_t x_t \right)^T \left( \theta - \sum_t \lambda_t y_t x_t \right) \right) d\theta \\
= \exp \left( \frac{1}{2} \left( \sum_t \lambda_t y_t x_t \right)^T \left( \sum_t \lambda_t y_t x_t \right) \right)
\]

As a result,

\[Z(\lambda) = \exp \left( \frac{1}{2} \left( \sum_t \lambda_t y_t x_t \right)^T \left( \sum_t \lambda_t y_t x_t \right) \right) \exp \left( \frac{1}{2} \sigma_b^2 \left( \sum_t \lambda_t y_t \right)^2 \right) \prod_t \frac{1}{1 - \lambda_t/c} \exp(-\lambda_t)\]

And we find the Lagrange multipliers \(\lambda^*\) by maximizing

\[\mathcal{J}(\theta) = -\frac{1}{2} \left( \sum_t \lambda_t y_t x_t \right)^T \left( \sum_t \lambda_t y_t x_t \right) - \frac{1}{2} \sigma_b^2 \left( \sum_t \lambda_t y_t \right)^2 + \sum_t (\log(1 - \lambda_t/c) + \lambda_t)\]

Once we have obtained the optimal \(\lambda^*\), we use these to generate a predictive label for \(x_{\text{new}}\). First we observe that

\[P(\theta) = \frac{1}{Z_\theta(\lambda)Z_b(\lambda)} P_0(\theta) \exp \left( \sum_t \lambda_t y_t \theta^T x_t \right) P_0(b) \exp \left( \sum_t \lambda_t y_t b \right)\]

where \(Z_\theta(\lambda) = \exp \left( \frac{1}{2} \left( \sum_t \lambda_t y_t x_t \right)^T \left( \sum_t \lambda_t y_t x_t \right) \right)\) and \(Z_b(\lambda) = \exp \left( \frac{1}{2} \sigma_b^2 \left( \sum_t \lambda_t y_t \right) \right)\)

Then,

\[\int P(\theta) \mathcal{L}(X|\theta) d\theta = \int P(\theta) (\theta^T x_{\text{new}} + b) d\theta = \int \frac{1}{Z_\theta(\lambda)} \theta^T x_{\text{new}} P_0(\theta) \exp \left( \sum_t \lambda_t y_t \theta^T x_t \right) d\theta + \int \frac{1}{Z_b(\lambda)} b P_0(b) \exp \left( \sum_t \lambda_t y_t b \right) db\]

where
Thus,\[
\hat{y} = \text{sign} \left( \sum_t \lambda_t y_t x_t \right)^T x_{\text{new}} + \sigma_b^2 \sum_t \lambda_t y_t
\]

We then implemented the above in \texttt{Matlab}. To maximize $J(\lambda)$ and solve for $\lambda$, we made use of \texttt{Matlab}'s existing functionality for solving constrained nonlinear problems. Once $\lambda^\star$ was obtained based on training data, we used the expression for $\hat{y}$ to classify test data. We now apply this implementation to the SONAR dataset and the Washington Wizards dataset mentioned in the Introduction.

### 3 Experiments

Not including toy experiments that generate data from two normal distributions with different means but the same covariance, our first experiment is the use of linear MED on the SONAR, Mines vs. Rocks dataset. Again, we use MED on this dataset to attempt to discriminate between sonar signals bounced off a metal cylinder (a “mine”) and those bounced off a roughly cylindrical rock. Quoting directly from the problem description on the UCI Machine Learning Repository, “The ”mines” [data] contains 111 patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. The ”rocks” [data] contains 97 patterns obtained from rocks under similar conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. The data set contains signals obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock.
Each pattern is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time. The integration aperture for higher frequencies occur later in time, since these frequencies are transmitted later during the chirp.”

To test our MED classifier on this dataset, we use the 10 times repeated 10-fold cross-validation version of this dataset used in “Benchmarking Support Vector Machines” by Meyer, Leisch & Hornik (2002). We then measure the average correct classification rate as the average across the different training/test sets of the percentage of test data points correctly classified. Our results are listed below in Table 1. As we can see, we tested the MED approach using a number of different values of \( c \). Recall \( c \) is a parameter in the prior for the margins, \( \gamma \), and is an upper bound for \( \lambda_t \) in the MED solution. We furthermore set \( \sigma^2_b = 100 \).

<table>
<thead>
<tr>
<th>( c )</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>76.6%</td>
<td>77.3%</td>
</tr>
<tr>
<td>3</td>
<td>77.8%</td>
<td>76.2%</td>
</tr>
<tr>
<td>5</td>
<td>77.3%</td>
<td>76.6%</td>
</tr>
<tr>
<td>10</td>
<td>76.2%</td>
<td>76.2%</td>
</tr>
<tr>
<td>20</td>
<td>74.8%</td>
<td>76.2%</td>
</tr>
<tr>
<td>100</td>
<td>73.9%</td>
<td>75.0%</td>
</tr>
</tbody>
</table>

Table 1: MED Avg. and Median Classification Rate for SONAR

Meyer, Leisch & Hornik referenced Gestel et. al 2001 in their review of previous results. Gestel et al. reported average accuracy of 73.3% for Linear Least-Squares SVM, 74.1% for Linear SVM and 67.9% for LDA for the SONAR dataset. Comparatively, using RBF SVM, Meyer, Leisch & Hornik report a mean misclassification rate of only 15.44% and find that a neural network procedure achieves a misclassification rate of only 10%. Veenman and Tax reported a 25.6% misclassification rate for linear SVM on the SONAR dataset although their cross-validation procedure is less clear. Nonetheless, it appears that our linear MED implementation compares favorably to reported linear SVM results for the SONAR dataset. This provides some confidence that our MED implementation is on the right path. Given the results of the RBF SVM and the neural net, naturally the results from a kernel or quadratic MED implementation would be of interest and we plan to implement this in future work. Also of interest would be a more refined procedure for choosing the parameter \( c \) as compared to our ad-hoc trial and error solution above.

We now turn to our second experiment involving the classification of win/loss data for the Washington Wizards from November 2003 to February 2008. Our inputs of interest are the point spread for the game, days of rest for the Wizards
leading up to the game, day of rest for the Wizards’ opponent leading up to the game and the over/under for the game. We have 400 games of data for this period. Over this period, the Wizards won 184 of those games and lost 216. The median point spread for these games was 2 with a min of point spread of -13 and max of 15. The median over/under with a min of 168.5 and a max 232. The median days of rest for the Wizards and their opponents was 1. This data was procurred from the www.sportsdatabase.com site. In Figure 1, we present a scatter plot of Over/Under vs. Point Spread for wins (“+”) and losses (“-”). From the plot, the association between point spread and win/loss outcome is clear (negative spreads indicate a team is favored). The association between over/under and outcome is less obvious, but if one exists wins seem to have a slight tendency to occur at higher over/under values while low over/under values seem more likely to be associated with losses for the Wizards.

![Figure 1: Scatterplot of Over/Under vs. Point Spread for Wizards Data](image)

We tested the MED linear classifier by selecting 200 games from the dataset as training data and using the remaining as test data. The was repeated 20 times with the same set of 200 games never repeated as training data (and hence a different set of 200 games always used for test data in each trial ). As above, we
set $\sigma^2_b = 100$. We set $c = 10$ based on ad-hoc trial and error. We compare these results to a naive classifier based on the direction of the point spread. If the MED Classifier outperforms the naive point spread classifier, this may suggest that the additional inputs contain information relevant to the win/loss outcome not reflected in the point spread.

The results of the trials are presented in Table 2. In general, MED appears to outperform the point spread but not significantly so. And, in Trial 6, MED drastically underperforms whereas the point spread classifier appears to encounter no such hiccup. The average classification rate for MED is 71.4 while the rate for naive point spread classification is 70.5. Thus, it is unclear whether there exists supplemental information value in the additional inputs.

<table>
<thead>
<tr>
<th>Trial</th>
<th>MED Classifier</th>
<th>Naive Point Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71.0%</td>
<td>68.5%</td>
</tr>
<tr>
<td>2</td>
<td>70.0%</td>
<td>68.0%</td>
</tr>
<tr>
<td>3</td>
<td>69.5%</td>
<td>67.5%</td>
</tr>
<tr>
<td>4</td>
<td>70.0%</td>
<td>68.5%</td>
</tr>
<tr>
<td>5</td>
<td>71.0%</td>
<td>70.0%</td>
</tr>
<tr>
<td>6</td>
<td>58.0%</td>
<td>71.0%</td>
</tr>
<tr>
<td>7</td>
<td>73.5%</td>
<td>72.5%</td>
</tr>
<tr>
<td>8</td>
<td>74.0%</td>
<td>72.5%</td>
</tr>
<tr>
<td>9</td>
<td>72.5%</td>
<td>71.0%</td>
</tr>
<tr>
<td>10</td>
<td>73.0%</td>
<td>71.5%</td>
</tr>
<tr>
<td>11</td>
<td>73.5%</td>
<td>72.5%</td>
</tr>
<tr>
<td>12</td>
<td>73.5%</td>
<td>73.0%</td>
</tr>
<tr>
<td>13</td>
<td>75.5%</td>
<td>73.5%</td>
</tr>
<tr>
<td>14</td>
<td>74.5%</td>
<td>72.5%</td>
</tr>
<tr>
<td>15</td>
<td>73.0%</td>
<td>71.0%</td>
</tr>
<tr>
<td>16</td>
<td>72.0%</td>
<td>70.0%</td>
</tr>
<tr>
<td>17</td>
<td>71.0%</td>
<td>68.5%</td>
</tr>
<tr>
<td>18</td>
<td>70.5%</td>
<td>68.5%</td>
</tr>
<tr>
<td>19</td>
<td>70.5%</td>
<td>70.0%</td>
</tr>
<tr>
<td>20</td>
<td>71.0%</td>
<td>69.5%</td>
</tr>
<tr>
<td>Average</td>
<td>71.4%</td>
<td>70.5%</td>
</tr>
</tbody>
</table>

Table 2: Win/Loss Classification Results for Wizards Data - Test Sets

As with the SONAR dataset, it would certainly be interesting to apply kernel or quadratic MED to this dataset. It would be of interest to compare these results to results obtained using other classifiers. Further expansion of this semi-manually collected dataset would also be of interest with more time. Finally, as above, it would be of interest to investigate both the choice of prior distribution for the margins and the choice of the parameter $c$. As the discussed above, we
have made somewhat adhoc choices for the distribution and parameter and we hope to focus future work on more refined choices in these areas.

References


