StatML Assignment2

Alexander Wahl-Rasmussen - LGV740
Maria Barrett - DTQ912
Keith Lia - DHL107

March 4, 2014

1 Classification

1.1 Linear Discriminant Analysis

Train Error	Test Error
0.15	0.2105

Table 1: Result of our own implementation of LDA

1.2 LDA and normalization

We normalized the train set to have zero mean and unit variance and transformed the test set using the mean and variance from the train set like in the previous assignment. If both train and test set are i.i.d this sets the mean to around 0 and the variance to around 1 in both sets. As seen from the result below, normalizing the data didn't change the performance.

The reason for that is, that the LDA assumes that the independent variables are normally distributed and that the covariance for all Gaussian distributions of the classes is the same without normalization. Therefore after normalization it is just the case and the result will not differ.

Train Error	Test Error
0.15	0.2105

Table 2: Result of our own implementation of LDA, normalized or transformed

1.3 Bayes optimal classification and probabilistic classification

Suppose we have a set of data, S, that we want to learn from. Specifically, what we want to learn is how to classify a new instance X with a label $y_k \in Y$. An intuitive approach could be to create a hypothesis h that in some way reflect the distribution S. However, it can prove challenging to find the optimal hypothesis - especially if you only have one hypothesis.

Instead it might prove fruitful to create a set of hypotheses $h_i \in H$ where each hypothesis is deterministic as to try to locate the optimal or most probably hypothesis. The optimal classification of a new instance is then obtained by first combining the predictions of all hypotheses, weighted by their posterior probabilities. The optimal hypothesis of the set is then the one with the lowest risk, i.e. the lowest amount of misclassifications, given we define the risk as being equal to the 0-1 loss . This method of finding the optimal hypothesis is also known as the Bayes optimal classification. Let us now see how we can use this method.

Formally we can define a classification problem as:

$$X = \{0\}, \qquad Y = \{0, 1\}, \qquad S = \{(0, 0), (0, 1), (0, 1), (0, 1)\}$$

Where the Bayes optimal classification for a set of hypotheses $h_i \in H$ is:

$$\text{BayesOpt} = \arg\max_{y_k \in Y} \sum_{h_i \in H} P(y_k|h_i)P(h_i|S)$$

We define the empirical risk as being equal to the loss of any given deterministic hypothesis h as the sum of the predictions h_i that does not correspond to the actual class y_k:

$$R_s(h) = \frac{1}{l} \sum_{i=1}^{l} \mathbb{I}(h(x_i) \neq y_k) = 1 - \sum_{h_i \in H} P(1|h_i)P(h_i|S)$$

We then define two deterministic hypotheses h_0 and h_1 that classifies every instance as either 0 or 1, where $H = \{h_0, h_1\}$. We then compute:

$$P(h_0|S) = \frac{1}{4} = 0.25, \qquad P(0|h_0) = 1, \qquad P(1|h_0) = 0$$
$$P(h_1|S) = \frac{3}{4} = 0.75, \qquad P(0|h_1) = 0, \qquad P(1|h_1) = 1$$

Thus the sum of each hypotheses probability of classifying S correctly is:

$$\sum_{y_k \in Y} P(0|h_i)P(h_i|S) = 1 * 0.25 + 0 * 0.75 = 0.25$$
$$\sum_{y_k \in Y} P(1|h_i)P(h_i|S) = 0 * 0.25 + 1 * 0.75 = 0.75$$

Which leaves us with the Bayes optimal classification as:

$$\arg\max_{y_k \in Y} \sum_{h_i \in H} P(y_k|h_i)P(h_i|S) = h_1$$

$$R_s(h_1) = \frac{1}{l} \sum_{i=1}^{l} \mathbb{I}(h_1(x_i) \neq y_k) = 0.25$$

If we instead assume that the hypotheses are in fact a single probabilistic hypothesis, then we view the prediction of h given an input X as a random variable and consider the probability of when $P(h(x) = y_k)$. We define the probabilistic classifier as predicting label 0 with probability of 0.25 and 1 with 0.75. Due to the relationship between the predicition accuracy and the risk of misclassification, and knowing the prediction accuracy, we can identify the loss function as:

$$L(y_k, y) = 1 - P(h(x) = y_k)$$

We can then write the expectation over the possible outcomes as:

$$R_s(h) = \sum_{y_k \in Y} P(h(x) = y_k) L(y_k, y)$$
$$R_s(h) = 0.25 * (1 - 0.25) + 0.75 * (1 - 0.75) = 0.375$$

We see then that the Bayes optimal classifier outperforms the probabilistic due to its risk being lower.

2 Regression: Sunspot Prediction

2.1 Maximum likelihood solution

To solve the Maximum Likelihood, we first create a design matrix for our variables. We use a linear basis function of the form $\phi_i(x) = x_i$, and $\phi_0(x) = 1$, so that it looks like the following:

$$\Phi = \begin{pmatrix} 1 & \phi_1(x1) & \cdots & \phi_{M-1}(x1) \\ 1 & \phi_1(x2) & \cdots & \phi_{M-1}(x2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(x_n) & \cdots & \phi_{M-1}(x_n) \end{pmatrix}$$

We then find the weight vector w_{ML} by applying the following formula:

$$w_{ML} = (\Phi^T \Phi)^{-1} \Phi^T t$$

For the subset D=1, we plot each x in the training set to its actual value, and each x in the test test to its actual value and predicted value, as seen in Figure 1. As seen in the plot, the predicted values for the training set have been regressed on a single line (black line is drawn for convenience).

Table 3 shows the Root Mean Square for each variable selection D=$\{1,2,5\}$. As we can see, taking all of the 5 variables provides the best result. Taking only the last variable (year s-16) gives a better score than taking variables 3 and 4 (year s-8 & year s-4). This means that the fifth variable gives a better indication of the amount of sunspots in the current year than the third and fourth variables

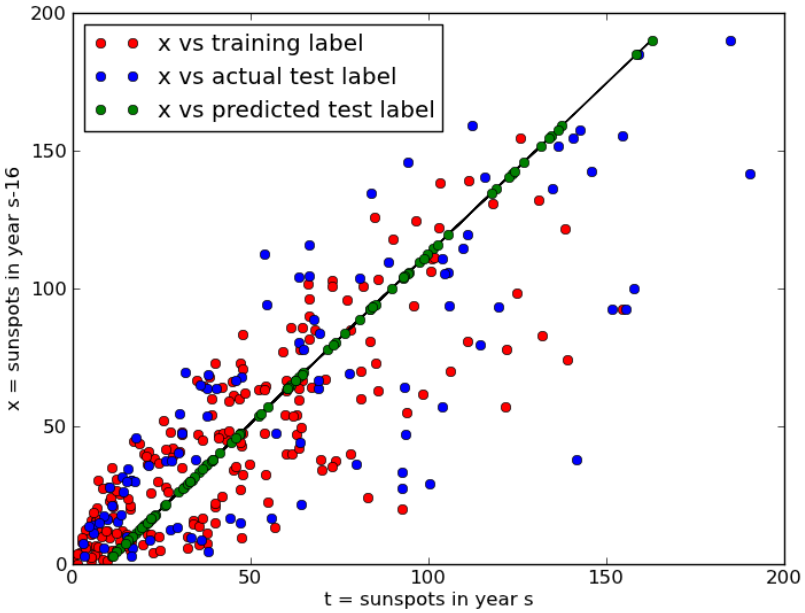

Figure 1: Plot of training and test datasets - actual and predicted

taken together. Figure 2 shows a plot of years (from 1916-2011) on the x-axis, versus the number of sunspots in the y-axis. We can see the actual amount in green, and the predicted results in red, blue and yellow.

Selection	Score
1	35.465059
2	28.839768
3	18.770007

Table 3: Root Mean Square Error for each variable selection

2.2 Maximum a posteriori solution

Our next task was to try the same exercise using Bayesion Maximum a posteriori. We used the formulae: $m_N = \beta S_N \Phi^T t$ for the mean and $S_N^{-1} = \alpha I + \beta \Phi^T \Phi$ for the covariance. We assigned the noise precision parameter β to 1, and set the range of α between 0 and 160, in increments of 5.

Figure 3 shows the Root Mean Square results for the different αs. From the results, when α=0, the RMS for variable selection 1 (D=2) is the same as Maximum Likelihood (RMS=35.46505899), but as we can see from the plot, as

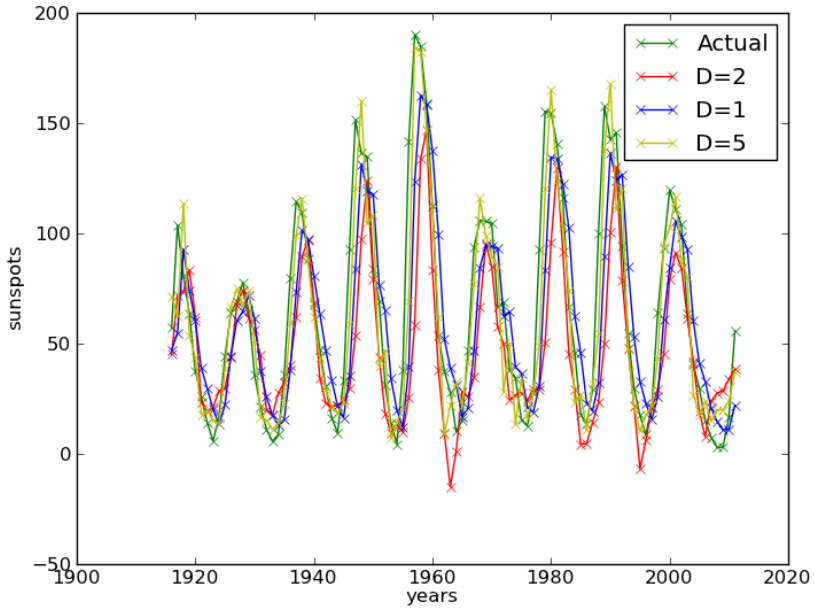

Figure 2: Plot of years vs number of sunspots

aplha increases, the RMS gets much worse, tending towards 47. For variable selection 2 (D=1), α=0 also gives the same result as ML (RMS=28.83976766), and as α increases, the score also gets worse, but at a much lower rate.

Furthermore, for the last selection (D=5), α=0 again gives the same result as ML (RMS=18.77000748), but as α gets bigger, the RMS score gets marginally lower. As α grows beyond 25, the RMS increases again, but still remains beneath the score for ML.

2.3 Weighted sum-of-squares

Minimizing $E_D(w)$, we get:

$$\frac{\delta}{\delta w_i}\frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - w^T\phi(x_n)\}^2 = 0$$

$$= \frac{2}{2}\sum_{n=1}^{N} r_n\{t_n - w^T\phi(x_n)\}.\phi(x_n) = 0$$

Since $\phi(\bar{x})^T = (\phi_0(x), \phi_1(x), \ldots, \phi_{M-1}(x))$, we get:

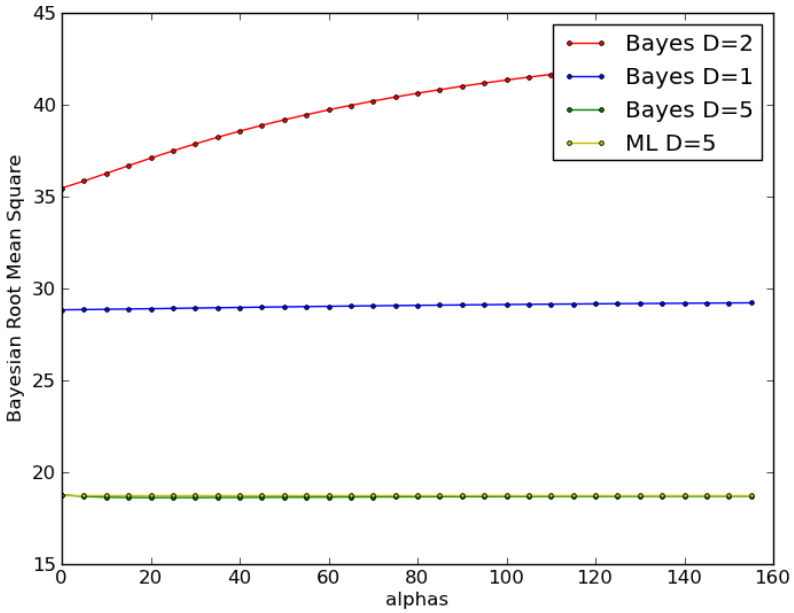

Figure 3: Plot of α vs RMS

$$\sum_{n=1}^{N} r_n t_n \phi(x_n)^T - \sum_{n=1}^{N} r_n w^T \phi(x_n)\phi(x_n)^T = 0$$

which can be re-written as:

$$rt^T\phi - w^T(\phi^T r\phi) = 0$$
$$=> w^T(\phi^T r\phi) = rt^T\phi$$
$$=> (\phi^T r\phi)w = (\phi^T r\phi)w = \phi^T rt$$
$$=> w^* = (\phi^T r\phi)^{-1}.\phi^T rt$$

2.3.1 Data dependent noise variance

If a variable has a high degree of variance, then it is intuitive to assume that that variable will have a low weighting. The variance describes how far away from the mean can a data point be. If there is low variance, then the data points are concentrated in a smaller space, and hence should hold more weight than a data point which is more spread out. In fact, the weighting terms **r** can be thought of as the inverse of the variance of the data point (x_n, t_n).

2.3.2 Replicated data points

If a data point is replicated in our data set, then it may bias our prediction by reinforcing the result of that instance. With the weighted sum-of-squares these replicated data points should be given a lower value to minimize this error rate.