THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

INTRODUCTION TO MACHINE LEARNING

DIT 45100

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Module 3

Linear Classification Techniques

Classification

- Categorical target feature
- Binary classification
- Multinomial classification
- Logistic regression
- Support Vector Machines (SVM)

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Logistic Regression

Logistic Regression

Scenario

- You work for a power utility company as an AI professional. The company owns a number of power generation stations, where electric generators operate day and night to serve clients uninterrupted. Condition of these generators is continuously monitored by measuring machine features indicative of their “health”.
- You are tasked to develop a classification model to predict the status of generators based on historic sensor measurements in order to avoid any unexpected machine breakdown.
- The objective is to classify generators as “good” or “faulty” based on these feature measurements

Generators dataset

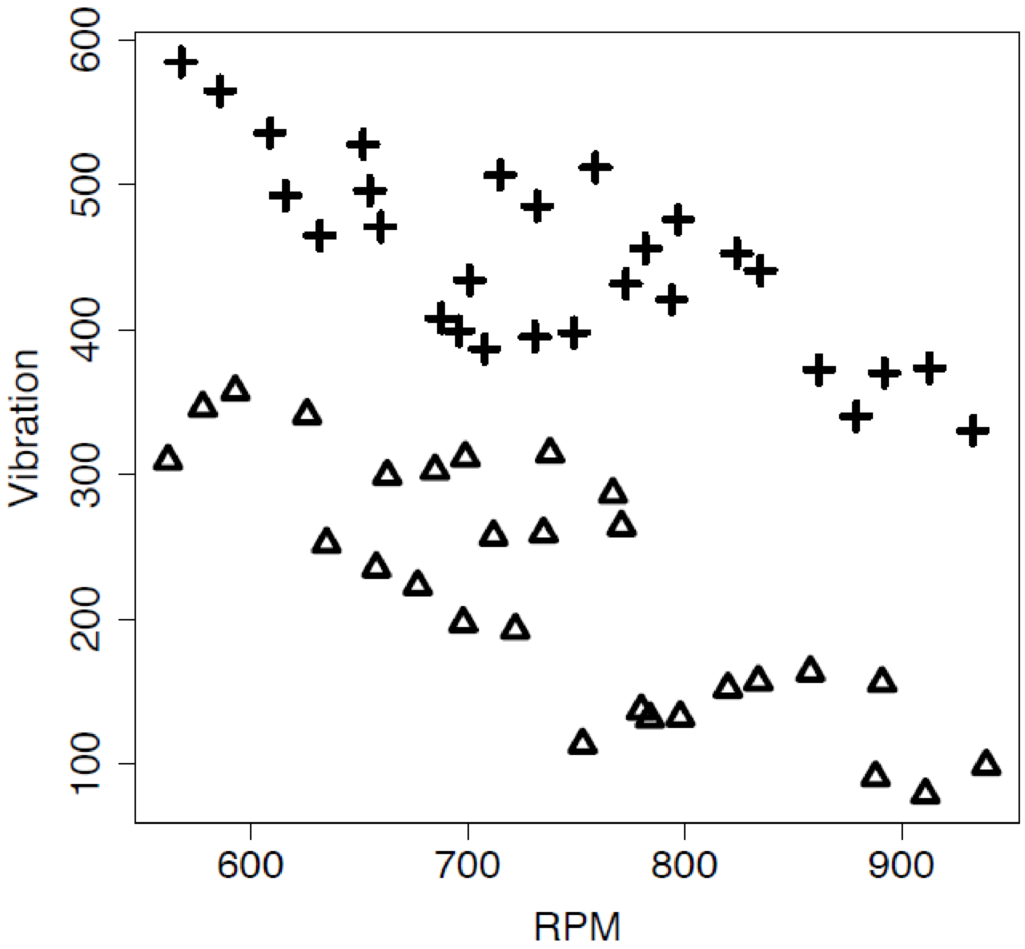
ID	RPM	VIBRATION	STATUS
1	568	585	good
2	586	565	good
3	609	536	good
4	616	492	good
5	632	465	good
6	652	528	good
7	655	496	good
8	660	471	good
9	688	408	good
10	696	399	good
11	708	387	good
12	701	434	good
13	715	506	good
14	732	485	good
15	731	395	good
16	749	398	good
17	759	512	good
18	773	431	good
19	782	456	good
20	797	476	good
21	794	421	good
22	824	452	good
23	835	441	good
24	862	372	good
25	879	340	good
26	892	370	good
27	913	373	good
28	933	330	good

ID	RPM	VIBRATION	STATUS
29	562	309	faulty
30	578	346	faulty
31	593	357	faulty
32	626	341	faulty
33	635	252	faulty
34	658	235	faulty
35	663	299	faulty
36	677	223	faulty
37	685	303	faulty
38	698	197	faulty
39	699	311	faulty
40	712	257	faulty
41	722	193	faulty
42	735	259	faulty
43	738	314	faulty
44	753	113	faulty
45	767	286	faulty
46	771	264	faulty
47	780	137	faulty
48	784	131	faulty
49	798	132	faulty
50	820	152	faulty
51	834	157	faulty
52	858	163	faulty
53	888	91	faulty
54	891	156	faulty
55	911	79	faulty
56	939	99	faulty

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

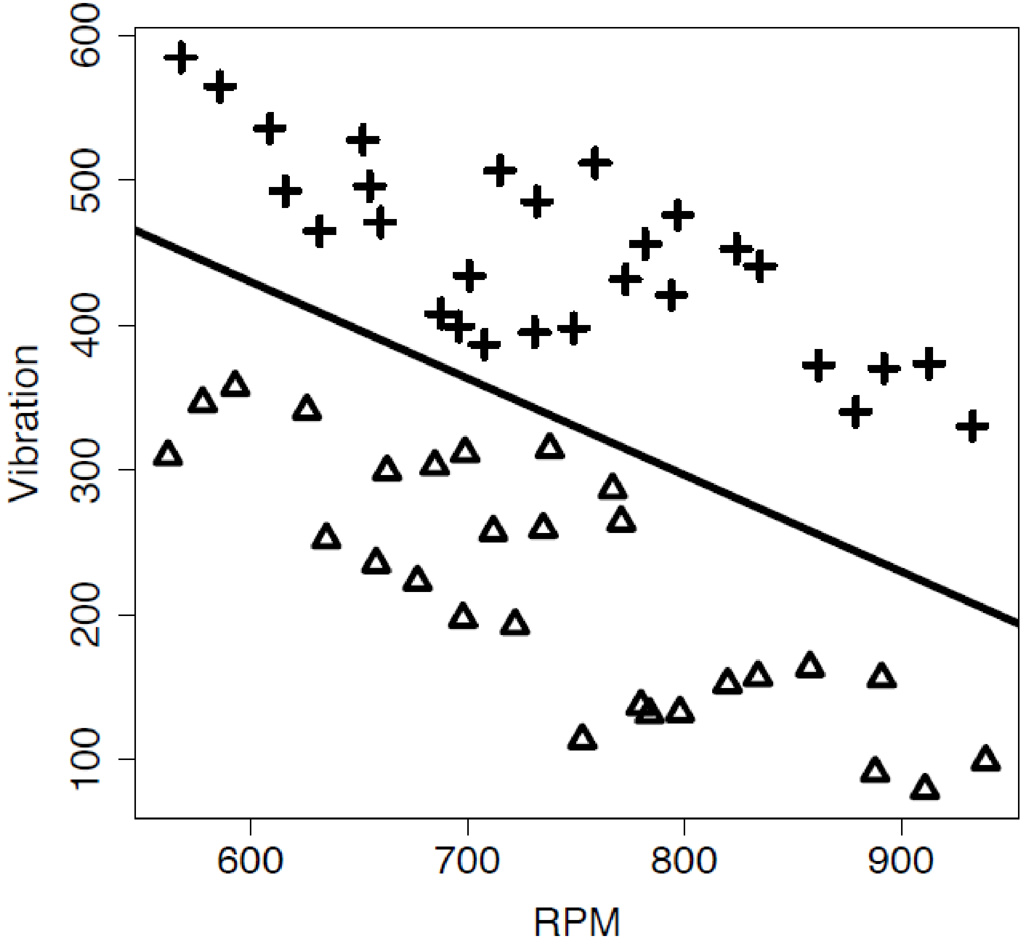
Generators dataset

Figure: A scatter plot of the RPM and VIBRATION descriptive features from the generators dataset, where ’good’ generators are shown as crosses and ’faulty’ generators are shown as triangles.



Generators dataset

Figure: A scatter plot of the RPM and VIBRATION descriptive features from the generators dataset, where ’good’ generators are shown as crosses and ’faulty’ generators are shown as triangles.



Logistic Regression

- As the decision boundary is a **linear separator** it can be defined using the equation of the line as:

$$\text{VIBRATION} = 830 - 0.667 \times \text{RPM} \tag{1}$$

or

$$830 - 0.667 \times \text{RPM} - \text{VIBRATION} = 0 \tag{2}$$

Logistic Regression

- Applying Equation (2) to the instance $\text{RPM} = 810$, $\text{VIBRATION} = 495$, which is **above** the decision boundary, gives the following result:

$$830 - 0.667 \times 810 - 495 = -205.27$$

- By contrast, if we apply Equation (2) to the instance $\text{RPM} = 650$ and $\text{VIBRATION} = 240$, which is **below** the decision boundary, we get

$$830 - 0.667 \times 650 - 240 = 156.45$$

Logistic Regression

- All the data points above the decision boundary will result in a negative value when plugged into the decision boundary equation,
- While all data points below the decision boundary will result in a positive value.

Logistic Regression

- Reverting to our previous notation we have:

$$\mathbb{M}_{\mathbf{w}}(\mathbf{d}) = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{d} \geq 0 \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

- The surface defined by this rule is known as a **decision surface**.

Logistic Regression

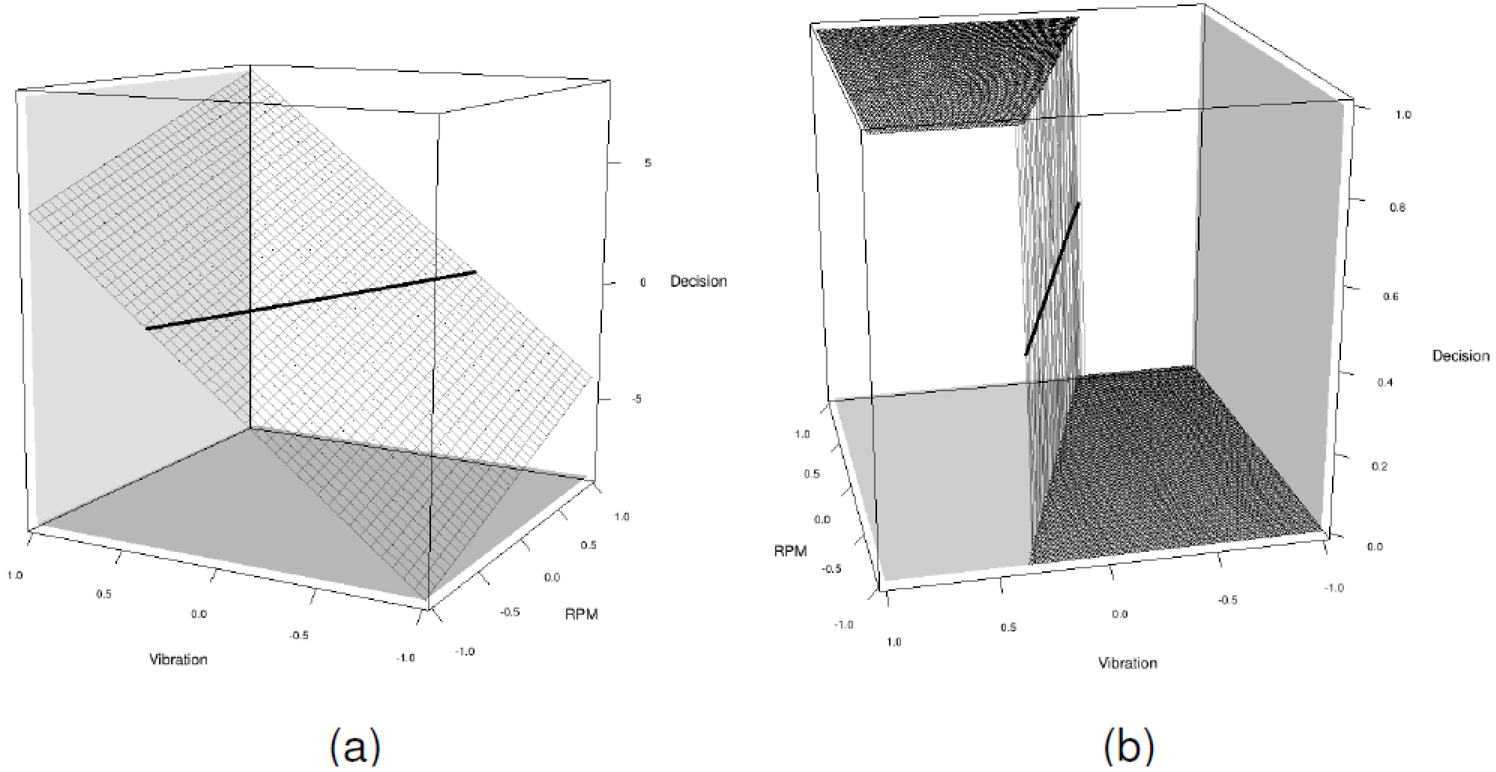


Figure: (a) A surface showing the value of Equation (2) for all values of RPM and VIBRATION. The decision boundary given in Equation (2) is highlighted. (b) The same surface linearly thresholded at zero to operate as a predictor.

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Logistic Regression

Logistic Regression

- The hard decision boundary given in Equation (3) is **discontinuous** so is not differentiable and so we can't calculate the gradient of the error surface.
- Furthermore, the model always makes completely confident predictions of 0 or 1, whereas a little more subtlety is desirable.
- We address these issues by using a more sophisticated threshold function that is continuous, and therefore differentiable, and that allows for the subtlety desired: the **logistic function**

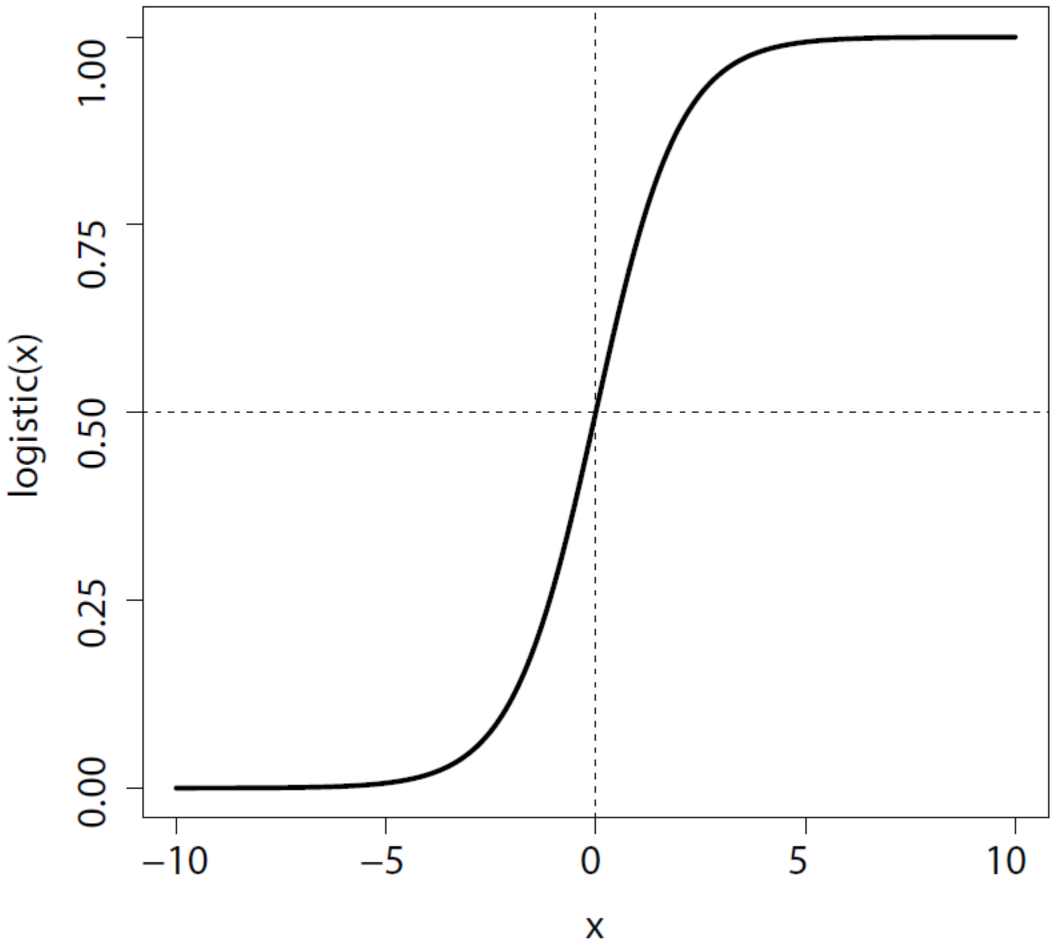
Logistic Regression

logistic function

$$Logistic(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

where x is a numeric value and e is **Euler's number** and is approximately equal to 2.7183.

Logistic Regression



Logistic Regression

- To build a logistic regression model, we simply pass the output of the basic linear regression model through the logistic function

$$\begin{aligned}\mathbb{M}_{\mathbf{w}}(\mathbf{d}) &= Logistic(\mathbf{w} \cdot \mathbf{d}) \\ &= \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{d}}}\end{aligned} \tag{5}$$

Logistic Regression

A note on training logistic regression models:

- Before we train a logistic regression model, we map the binary target feature levels to 0 or 1.
- The error of the model on each instance is then the difference between the target feature (0 or 1) and the value of the prediction [0, 1].

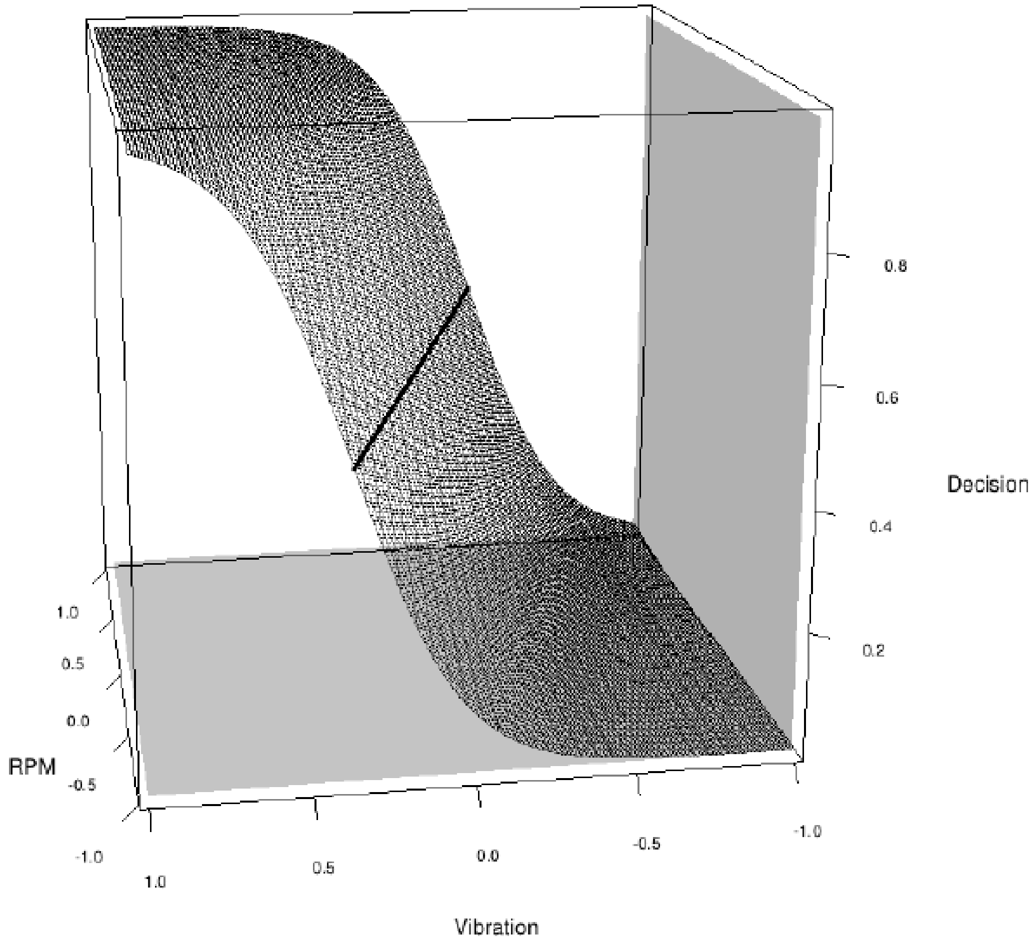
Logistic Regression

Example

$$\mathbb{M}_{\mathbf{w}}(\langle \text{RPM}, \text{VIBRATION} \rangle) = \frac{1}{1 + e^{-(-0.4077 + 4.1697 \times \text{RPM} + 6.0460 \times \text{VIBRATION})}}$$

Logistic Regression

Figure: The decision surface for the example logistic regression model.



Logistic Regression

$$P(t = \text{'faulty'} | \mathbf{d}) = \mathbb{M}_{\mathbf{w}}(\mathbf{d})$$

$$P(t = \text{'good'} | \mathbf{d}) = 1 - \mathbb{M}_{\mathbf{w}}(\mathbf{d})$$

Logistic Regression

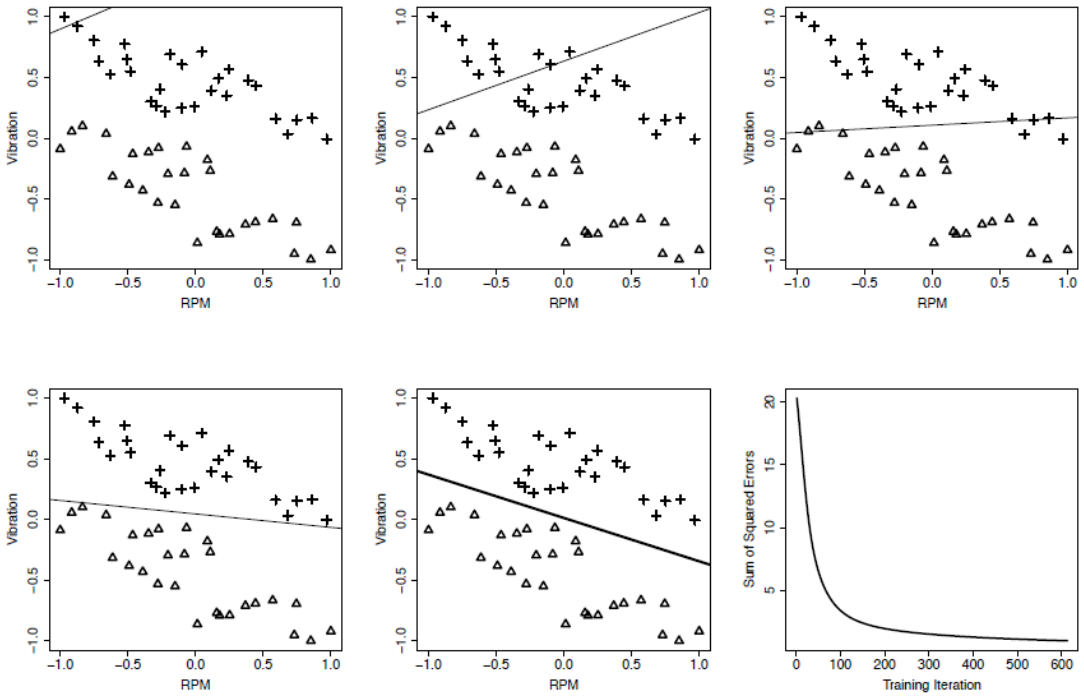


Figure: A selection of the logistic regression models developed during the gradient descent process for the machinery dataset. The bottom-right panel shows the sum of squared error values generated during the gradient descent process.

Logistic Regression

- To repurpose the gradient descent algorithm for training logistic regression models the only change that needs to be made is in the weight update rule.

- The new weight update rule is:

$$\mathbf{w}[j] \leftarrow \mathbf{w}[j] + \alpha \times \sum_{i=1}^{n} ((t - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i)) \times \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i) \times (1 - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i)) \times \mathbf{d}_i[j])$$

Logistic Regression

ID	RPM	VIBRATION	STATUS
1	498	604	faulty
2	517	594	faulty
3	541	574	faulty
4	555	587	faulty
5	572	537	faulty
6	600	553	faulty
7	621	482	faulty
8	632	539	faulty
9	656	476	faulty
10	653	554	faulty
11	679	516	faulty
12	688	524	faulty
13	684	450	faulty
14	699	512	faulty
15	703	505	faulty
16	717	377	faulty
17	740	377	faulty
18	749	501	faulty
19	756	492	faulty
20	752	381	faulty
21	762	508	faulty
22	781	474	faulty
23	781	480	faulty
24	804	460	faulty
25	828	346	faulty
26	830	366	faulty
27	864	344	faulty
28	882	403	faulty
29	891	338	faulty
30	921	362	faulty
31	941	301	faulty
32	965	336	faulty
33	976	297	faulty
34	994	287	faulty

ID	RPM	VIBRATION	STATUS
35	501	463	good
36	526	443	good
37	536	412	good
38	564	394	good
39	584	398	good
40	602	398	good
41	610	428	good
42	638	389	good
43	652	394	good
44	659	336	good
45	662	364	good
46	672	308	good
47	691	248	good
48	694	401	good
49	718	313	good
50	720	410	good
51	723	389	good
52	744	227	good
53	741	397	good
54	770	200	good
55	764	370	good
56	790	248	good
57	786	344	good
58	792	290	good
59	818	268	good
60	845	232	good
61	867	195	good
62	878	168	good
63	895	218	good
64	916	221	good
65	950	156	good
66	956	174	good
67	973	134	good
68	1002	121	good

Logistic Regression

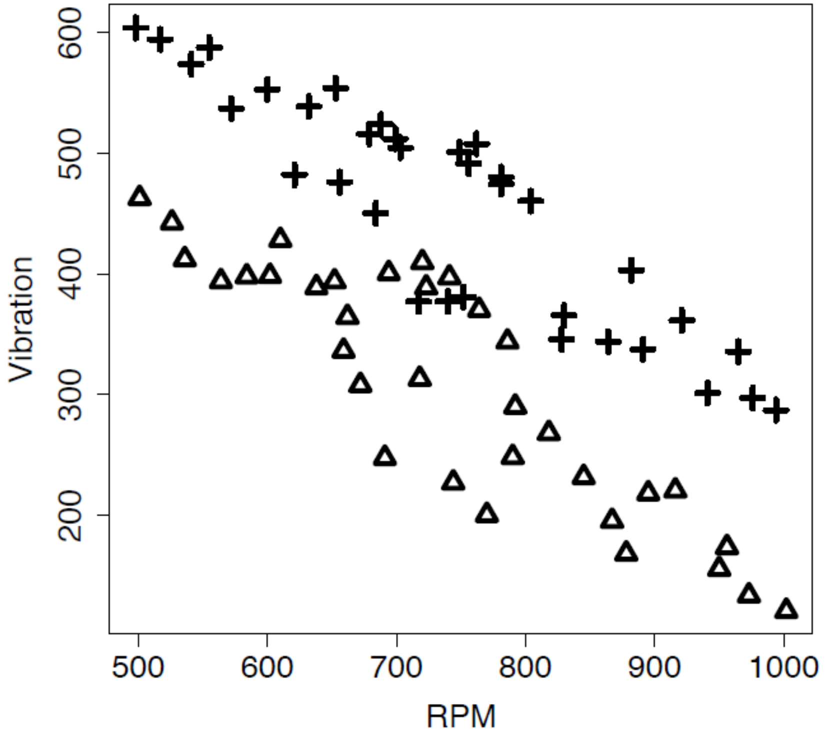


Figure: A scatter plot of the extended generators dataset, which results in instances with the different target levels overlapping with each other. ’good’ generators are shown as crosses, and ’faulty’ generators are shown as triangles.

Logistic Regression

For logistic regression models we recommend that descriptive feature values always be normalized.

In this example, before the training process begins, both descriptive features are normalized to the range [-1, 1].

Logistic Regression

For this example let's assume that:

- Learning rate:
  - $\alpha = 0.02$
- Initial Weights:
  - $\mathbf{w}[0] = -2.9465$
  - $\mathbf{w}[1] = -1.0147$
  - $\mathbf{w}[2] = 2.1610$

Logistic Regression

Iteration 1

ID	TARGET LEVEL	Pred.	Error	Squared Error	errorDelta($\mathcal{D}$, **w**[**i**]) **w**[0]	**w**[1]	**w**[2]
1	1	0.5570	0.4430	0.1963	0.1093	-0.1093	0.1093
2	1	0.5168	0.4832	0.2335	0.1207	-0.1116	0.1159
3	1	0.4469	0.5531	0.3059	0.1367	-0.1134	0.1197
4	1	0.4629	0.5371	0.2885	0.1335	-0.1033	0.1244
...							
65	0	0.0037	-0.0037	0.0000	0.0000	0.0000	0.0000
66	0	0.0042	-0.0042	0.0000	0.0000	0.0000	0.0000
67	0	0.0028	-0.0028	0.0000	0.0000	0.0000	0.0000
68	0	0.0022	-0.0022	0.0000	0.0000	0.0000	0.0000
			Sum	24.4738	2.7031	-0.7015	1.6493
Sum of squared errors (Sum/2)				12.2369			

Logistic Regression

$$\mathbf{w}[j] \leftarrow \mathbf{w}[j] + \alpha \times \sum_{i=1}^{n} ((t_i - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i)) \times \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i) \times (1 - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i)) \times \mathbf{d}_i[j])$$

New Weights (after Iteration 1)		
w[0] = -2.8924	**w**[1] = -1.0287	**w**[2] = 2.1940

Logistic Regression

Iteration 2

ID	Target Level	Pred.	Error	Squared Error	errorDelta($\mathcal{D}$, **w**[i]) **w**[0]	**w**[1]	**w**[2]
1	1	0.5817	0.4183	0.1749	0.1018	-0.1018	0.1018
2	1	0.5414	0.4586	0.2103	0.1139	-0.1053	0.1094
3	1	0.4704	0.5296	0.2805	0.1319	-0.1094	0.1155
4	1	0.4867	0.5133	0.2635	0.1282	-0.0992	0.1194
...							
65	0	0.0037	-0.0037	0.0000	0.0000	0.0000	0.0000
66	0	0.0043	-0.0043	0.0000	0.0000	0.0000	0.0000
67	0	0.0028	-0.0028	0.0000	0.0000	0.0000	0.0000
68	0	0.0022	-0.0022	0.0000	0.0000	0.0000	0.0000
			Sum	24.0524	2.7236	-0.6646	1.6484
			Sum of squared errors (Sum/2)	12.0262			

Logistic Regression

$$\mathbf{w}[j] \leftarrow \mathbf{w}[j] + \alpha \times \sum_{i=1}^{n} ((t_i - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i)) \times \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i) \times (1 - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i)) \times \mathbf{d}_i[j])$$

New Weights (after Iteration 2)		
w[0] = -2.8380	**w**[1] = -1.0416	**w**[2] = 2.2271

Logistic Regression

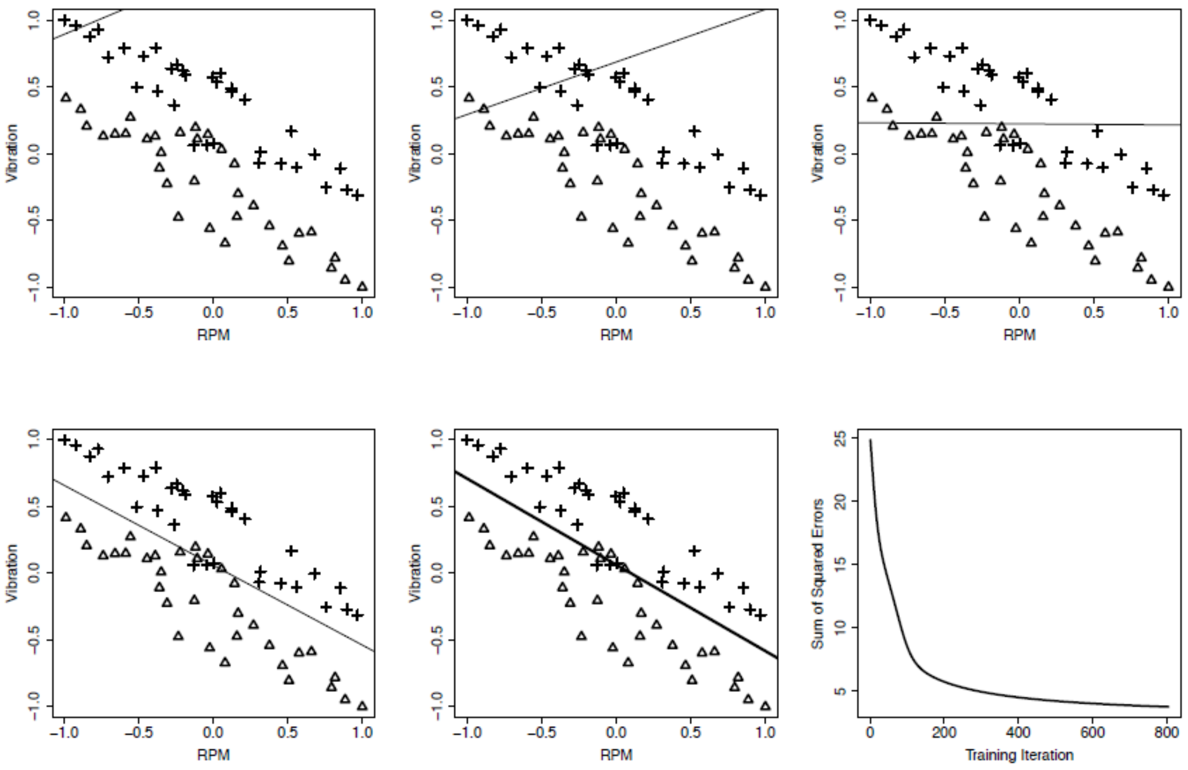


Figure: A selection of the logistic regression models developed during the gradient descent process for the extended generators dataset. The bottom-right panel shows the sum of squared error values generated during the gradient descent process.

Logistic Regression

- The final model found is:

$$\mathbb{M}_{\mathbf{w}}(\langle \text{RPM}, \text{VIBRATION} \rangle) = \frac{1}{1 + e^{-(-0.4077 + 4.1697 \times \text{RPM} + 6.0460 \times \text{VIBRATION})}}$$

Performance Measures

- Accuracy
- Confusion matrix
- Recall
- Precision
- F1 Score
- …

Confusion Matrix

CLASS		Predicted	
		Positive	**Negative**
Actual	**Positive**	TP	FN
	Negative	FP	TN

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Percision = \frac{TP}{TP + FP}$$

$$F1\ Score = \frac{2 \text{ x Precision x Recall}}{Precision + Recall}$$

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Modeling Non-Linear Relationships

Non-Linear Relationships

Table: A dataset showing participants' responses to viewing 'positive' and 'negative' images measured on the EEG P20 and P45 potentials.

ID	P20	P45	TYPE	ID	P20	P45	TYPE
1	0.4497	0.4499	negative	26	0.0656	0.2244	positive
2	0.8964	0.9006	negative	27	0.6336	0.2312	positive
3	0.6952	0.3760	negative	28	0.4453	0.4052	positive
4	0.1769	0.7050	negative	29	0.9998	0.8493	positive
5	0.6904	0.4505	negative	30	0.9027	0.6080	positive
6	0.7794	0.9190	negative	31	0.3319	0.1473	positive
	⋮				⋮		

Non-Linear Relationships

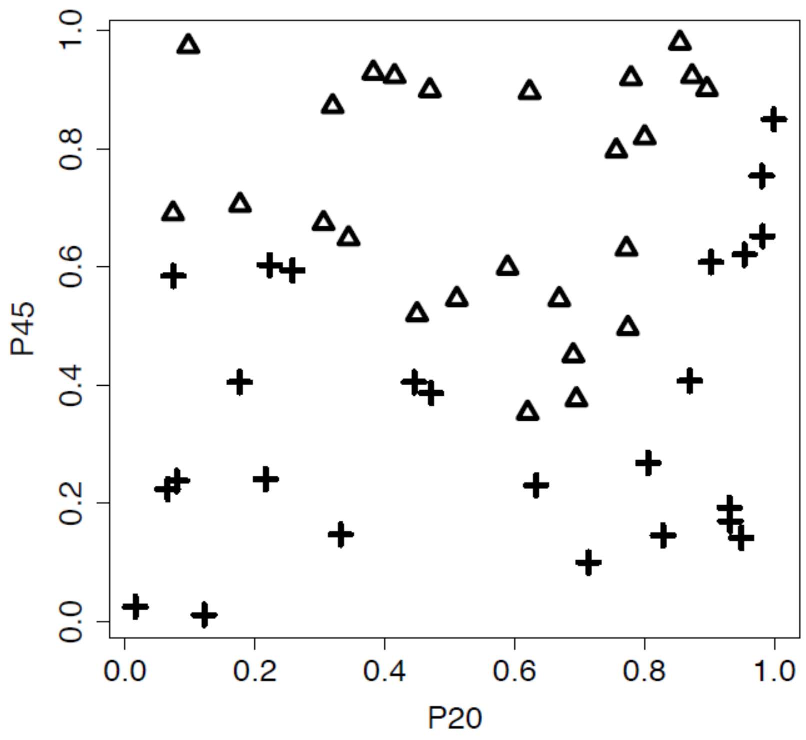


Figure: A scatter plot of the P20 and P45 features from the EEG dataset. *'positive'* images are shown as crosses, and *'negative'* images are shown as triangles.

Non-Linear Relationships

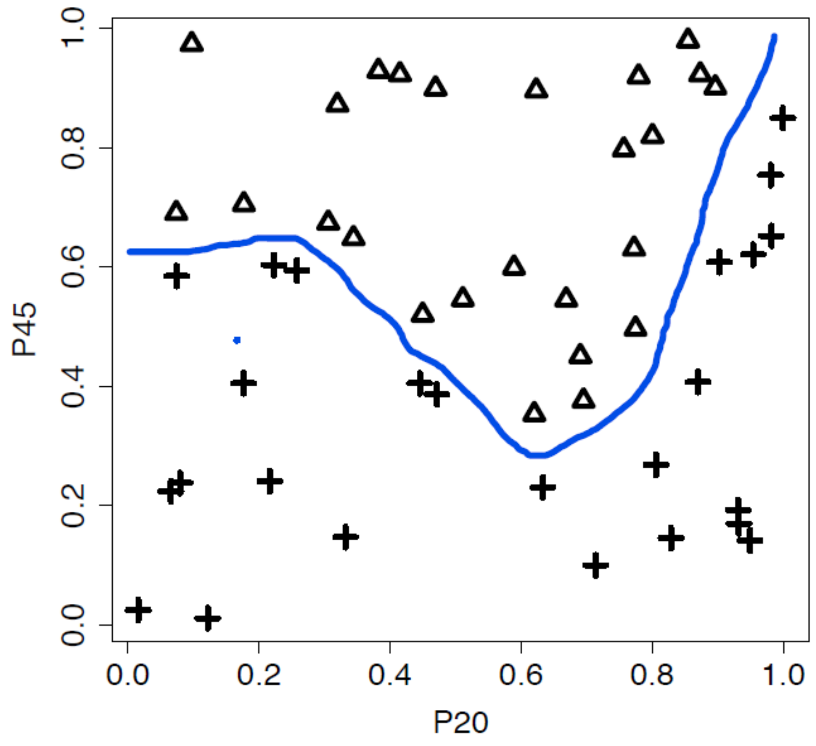


Figure: A scatter plot of the P20 and P45 features from the EEG dataset. *'positive'* images are shown as crosses, and *'negative'* images are shown as triangles.

Non-Linear Relationships

- A logistic regression model using basis functions is defined as follows:

$$\mathbb{M}_{\mathbf{w}}(\mathbf{d}) = \frac{1}{1 + e^{-\left(\sum_{j=0}^{b} \mathbf{w}[j]\phi_j(\mathbf{d})\right)}} \quad (6)$$

Non-Linear Relationships

- We will use the following basis functions for the EEG problem:

$$\begin{aligned} \phi_0(\langle \text{P20}, \text{P45} \rangle) &= 1 & \phi_4(\langle \text{P20}, \text{P45} \rangle) &= \text{P45}^2 \\ \phi_1(\langle \text{P20}, \text{P45} \rangle) &= \text{P20} & \phi_5(\langle \text{P20}, \text{P45} \rangle) &= \text{P20}^3 \\ \phi_2(\langle \text{P20}, \text{P45} \rangle) &= \text{P45} & \phi_6(\langle \text{P20}, \text{P45} \rangle) &= \text{P45}^3 \\ \phi_3(\langle \text{P20}, \text{P45} \rangle) &= \text{P20}^2 & \phi_7(\langle \text{P20}, \text{P45} \rangle) &= \text{P20} \times \text{P45} \end{aligned}$$

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Non-Linear Relationships

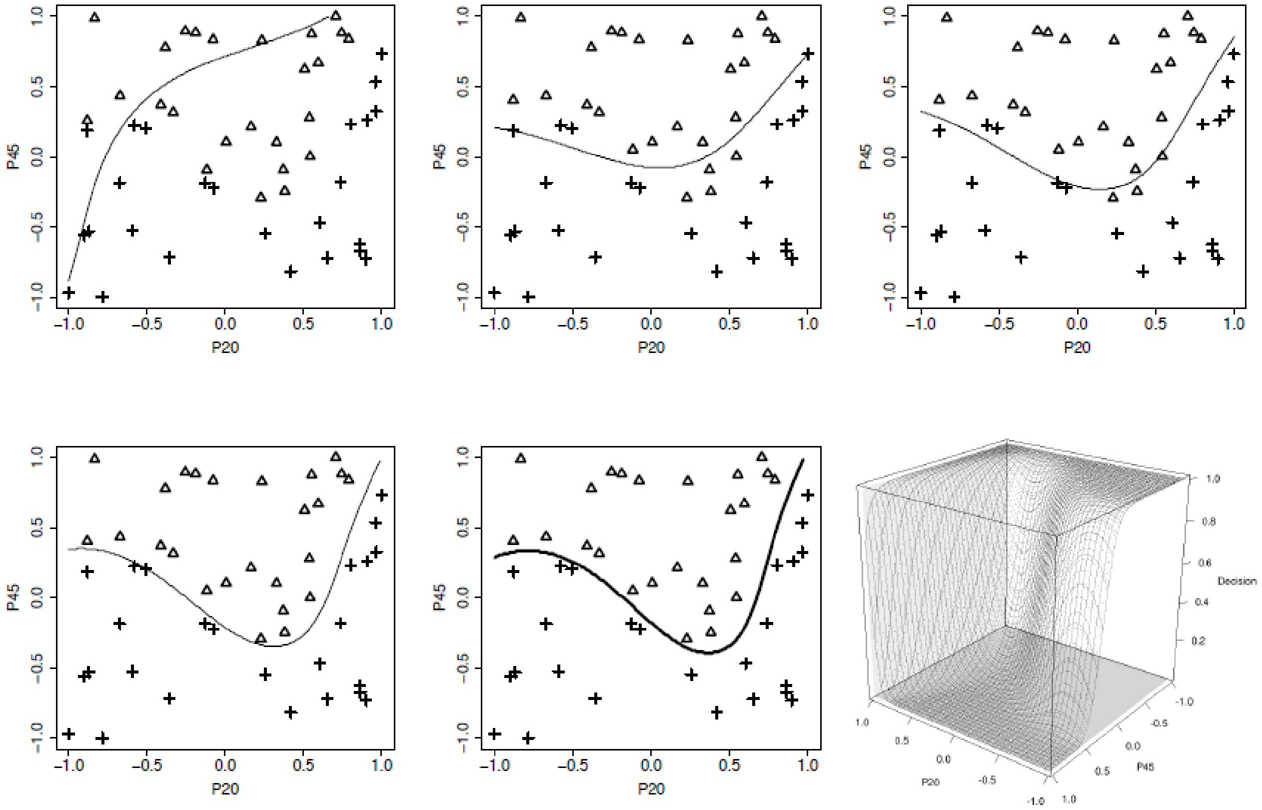


Figure: A selection of the models developed during the gradient descent process for the EEG dataset. The final panel shows the decision surface generated.