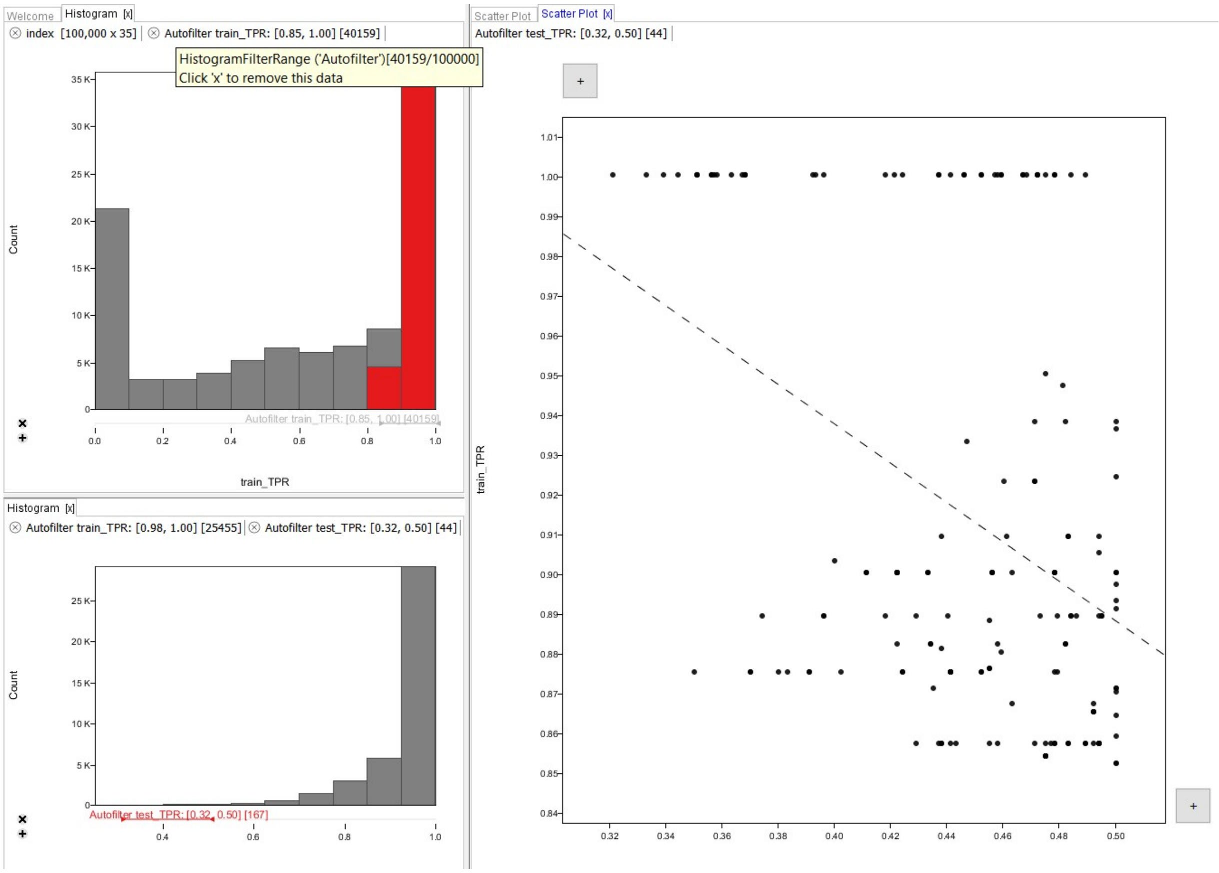
Overfitted Data

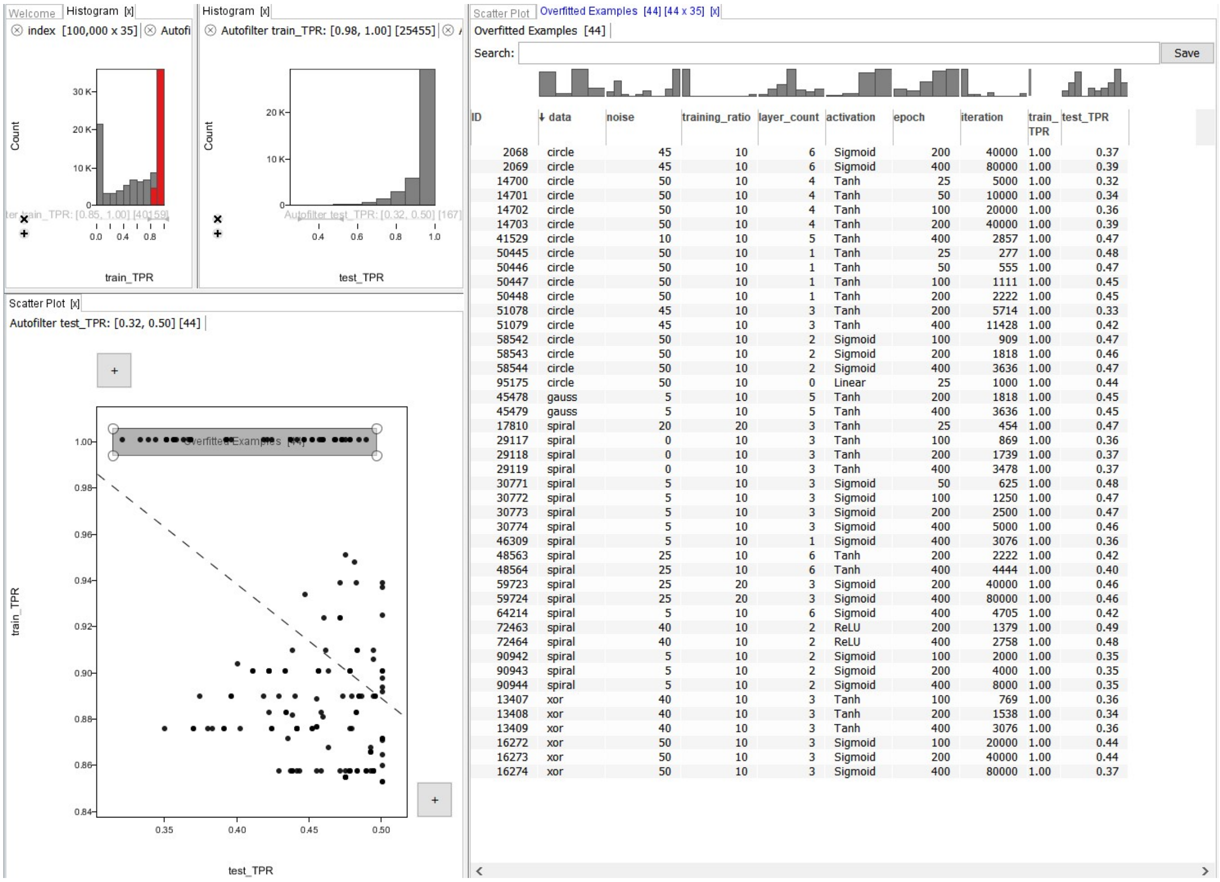
Notebook:	OstravaSummerSchool		
Created:	9/7/2017 07:01 PM	Updated:	9/7/2017 10:22 PM
Author:	manfred.klaffenboeck@univie.ac.at		

The original idea


HistogramFilterRange ('Autofilter')[40159/100000]
Click 'x' to remove this data

The true positive rate for the training data was set to 0.85. For the resulting dataset, the true positive rate for the test data was set to a threshold below 0.5. This only leaves a couple hundred results (167, to be exact). The expectation was, that this would at least correlate, meaning, that it happens less for a train_TPR with a high value (1.00) then for a lower one. As the scatterplot shows though, this is not accurate. A little research shows though, that this is a phenomenon known as overfitting.

Digging deeper



ID	↓ data	noise	training_ratio	layer_count	activation	epoch	iteration	train_ TPR	test_TPR
2068	circle	45	10	6	Sigmoid	200	40000	1.00	0.37
2069	circle	45	10	6	Sigmoid	400	80000	1.00	0.39
14700	circle	50	10	4	Tanh	25	5000	1.00	0.32
14701	circle	50	10	4	Tanh	50	10000	1.00	0.34
14702	circle	50	10	4	Tanh	100	20000	1.00	0.36
14703	circle	50	10	4	Tanh	200	40000	1.00	0.39
41529	circle	10	10	5	Tanh	400	2857	1.00	0.47
50445	circle	50	10	1	Tanh	25	277	1.00	0.48
50446	circle	50	10	1	Tanh	50	555	1.00	0.47
50447	circle	50	10	1	Tanh	100	1111	1.00	0.45
50448	circle	50	10	1	Tanh	200	2222	1.00	0.45
51078	circle	45	10	3	Tanh	200	5714	1.00	0.33
51079	circle	45	10	3	Tanh	400	11428	1.00	0.42
58542	circle	50	10	2	Sigmoid	100	909	1.00	0.47
58543	circle	50	10	2	Sigmoid	200	1818	1.00	0.46
58544	circle	50	10	2	Sigmoid	400	3636	1.00	0.47
95175	circle	50	10	0	Linear	25	1000	1.00	0.44
45478	gauss	5	10	5	Tanh	200	1818	1.00	0.45
45479	gauss	5	10	5	Tanh	400	3636	1.00	0.45
17810	spiral	20	20	3	Tanh	25	454	1.00	0.47
29117	spiral	0	10	3	Tanh	100	869	1.00	0.36
29118	spiral	0	10	3	Tanh	200	1739	1.00	0.37
29119	spiral	0	10	3	Tanh	400	3478	1.00	0.37
30771	spiral	5	10	3	Sigmoid	50	625	1.00	0.48
30772	spiral	5	10	3	Sigmoid	100	1250	1.00	0.47
30773	spiral	5	10	3	Sigmoid	200	2500	1.00	0.47
30774	spiral	5	10	3	Sigmoid	400	5000	1.00	0.46
46309	spiral	5	10	1	Sigmoid	400	3076	1.00	0.36
48563	spiral	25	10	6	Tanh	200	2222	1.00	0.42
48564	spiral	25	10	6	Tanh	400	4444	1.00	0.40
59723	spiral	25	20	3	Sigmoid	200	40000	1.00	0.46
59724	spiral	25	20	3	Sigmoid	400	80000	1.00	0.46
64214	spiral	5	10	6	Sigmoid	400	4705	1.00	0.42
72463	spiral	40	10	2	ReLU	200	1379	1.00	0.49
72464	spiral	40	10	2	ReLU	400	2758	1.00	0.48
90942	spiral	5	10	2	Sigmoid	100	2000	1.00	0.35
90943	spiral	5	10	2	Sigmoid	200	4000	1.00	0.35
90944	spiral	5	10	2	Sigmoid	400	8000	1.00	0.35
13407	xor	40	10	3	Tanh	100	769	1.00	0.36
13408	xor	40	10	3	Tanh	200	1538	1.00	0.34
13409	xor	40	10	3	Tanh	400	3076	1.00	0.36
16272	xor	50	10	3	Sigmoid	100	20000	1.00	0.44
16273	xor	50	10	3	Sigmoid	200	40000	1.00	0.44
16274	xor	50	10	3	Sigmoid	400	80000	1.00	0.37

Focusing only on the overfitted examples and putting the results in a table view, it can be very easily seen, that this phenomenon only occurs

These findings are somewhat expectable as well as reassuring in neural networks in general. At the same time though, it actually is interesting that the training ratio seems to be the only real influental factor in this matter. As can also be very easily seen through the histograms on top of the table view, it happens on all datatypes (although more often in spirals and circles than in the other datatypes), on various noise levels, layer counts, activation functions and epochs. In the case of the epochs, it seems interesting that apparently this seems to happen less often with less epochs.

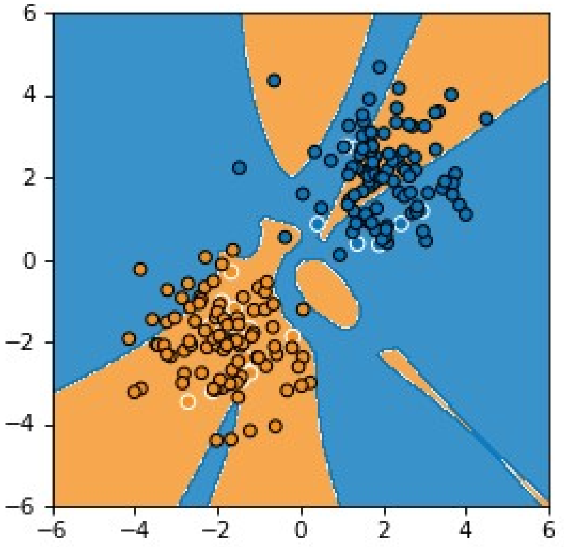
Example images

The following section showcases some of learned example images.

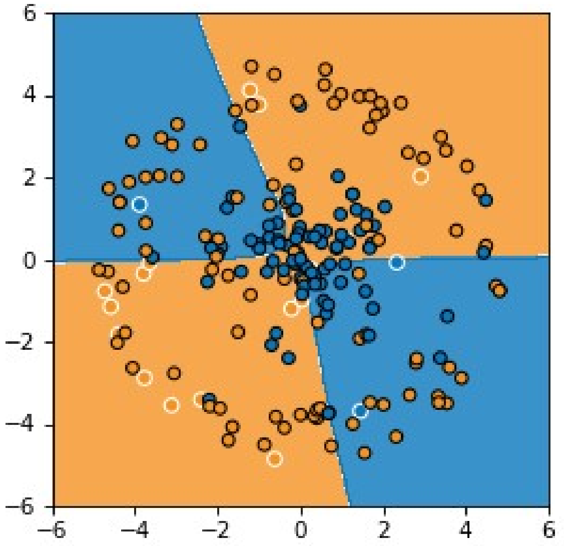
Dots with white circles around them mark the training data, dots with black circles mark the test data.

As can be seen, the background as learned by the NN always matches the dots with the white circles around them but does not nearly approximate the (more or less) obvious data clusters.

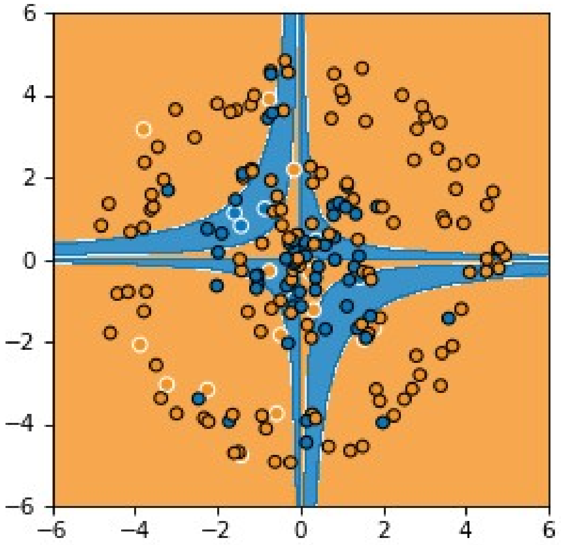
ID: 45478



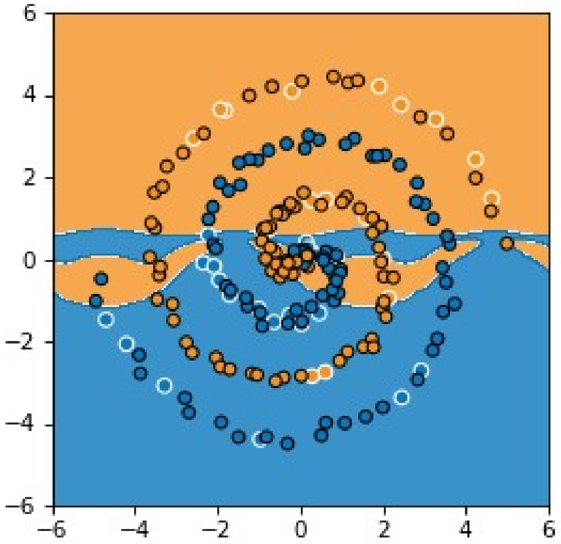
ID: 2068



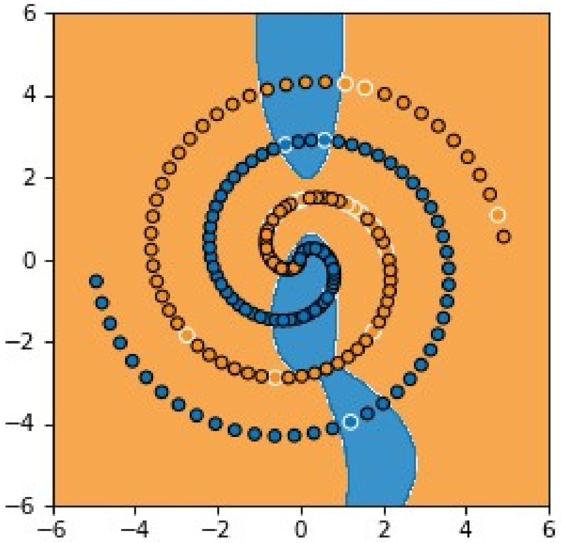
ID: 14700



ID: 17810



ID: 29119


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6