ML Final Project

Authors: Leo Li, Oliver Li

Introduction

- Stroke prediction dataset
  - Binary classification problem
- Models
  - Logistic
  - Support Vector Machine (SVM)
  - Neural Networks (NN)

Data

id	gender	age	hypertensi...	heart_dise...	ever_marri...	work_type	Residence...	avg_gluco...	bmi	smoking_s...	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1

Data Preparation

- Removed outliers
- Repaired incomplete features
- Categorical Encoding
  - Label Encoding
  - One-hot Encoding
- Normalization Scaling
- Data splits: 60% train, 20% validation, 20% test

Logistic Regression

1. Polynomial Feature Transformation
2. K-fold cross validation for regularization C selection

Polynomial Feature Transformation

Degree 1: 18 features

Degree 2: 171 features

Degree 3: 1140 features

Dataset	**Polynomial Transformation Degree**	
	1 (no transformation)	***2***
Train	0.95106	0.953997
Validation	0.962818	0.954012

Table 1: Polynomial transformation results

K-Fold CV with Regularization

C = λ^{-1}, 10 values tested from 0.0001 to 10000

L1 Regularization		
K	**Best C**	**Score (Accuracy)**
2	0.0001	0.95106
3	0.0001	0.95106
4	0.0001	0.95106
5	0.35938	0.95139
6	0.35938	0.95139
7	2.78256	0.95139
8	2.78256	0.95139
9	2.78256	0.95139
10	0.0001	0.95106

Table 2: Logistic L1 Regularization Results

K-Fold CV with Regularization

C = λ^{-1}, 10 values tested from 0.0001 to 10000

L2 Regularization		
K	**Best C**	**Score (Accuracy)**
2	0.0001	0.95106
3	0.0001	0.95106
4	0.0001	0.95106
5	0.35938	0.95139
6	0.35938	0.95139
7	0.35938	0.95139
8	0.35938	0.95139
9	2.78256	0.95139
10	0.0001	0.95106

Table 3: Logistic L2 Regularization results

Logistic Regression - Conclusion

Best model: No polynomial feature transformation, no regularization

Test set accuracy: 0.9413

Insights:

1. Best C values suggest data is very linearly separable
2. Rest of loss probably caused by irreducible noise

Support Vector Machine: Linear Kernel

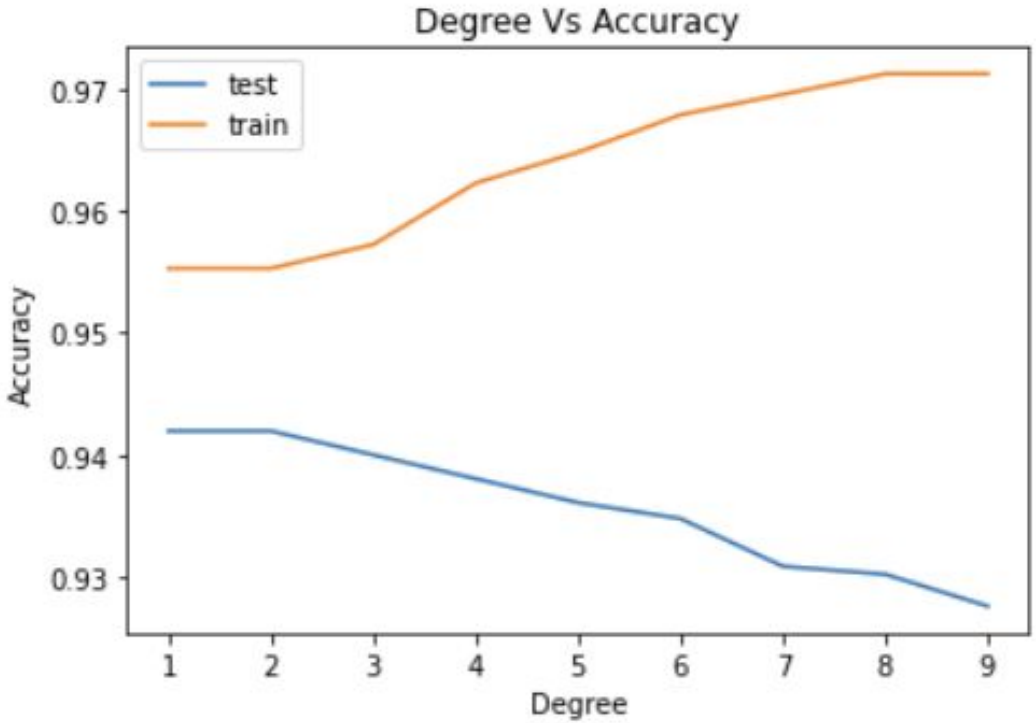
- Observed convergence of results

Dataset	Configuration		
	No Regularization	*L1 Norm*	*L2 Norm*
Train	0.955257	0.955257	0.955257
Test	0.941944	0.941944	0.941944

Support Vector Machine: Polynomial Kernel

Degree	Training	Test
1	0.955257	0.941944
2	0.955257	0.941944
3	0.957215	0.939987
4	0.962248	0.938030
5	0.964765	0.936073
6	0.967841	0.934768
7	0.969519	0.930855
8	0.971197	0.930202
9	0.971197	0.927593

Support Vector Machine: Polynomial Kernel



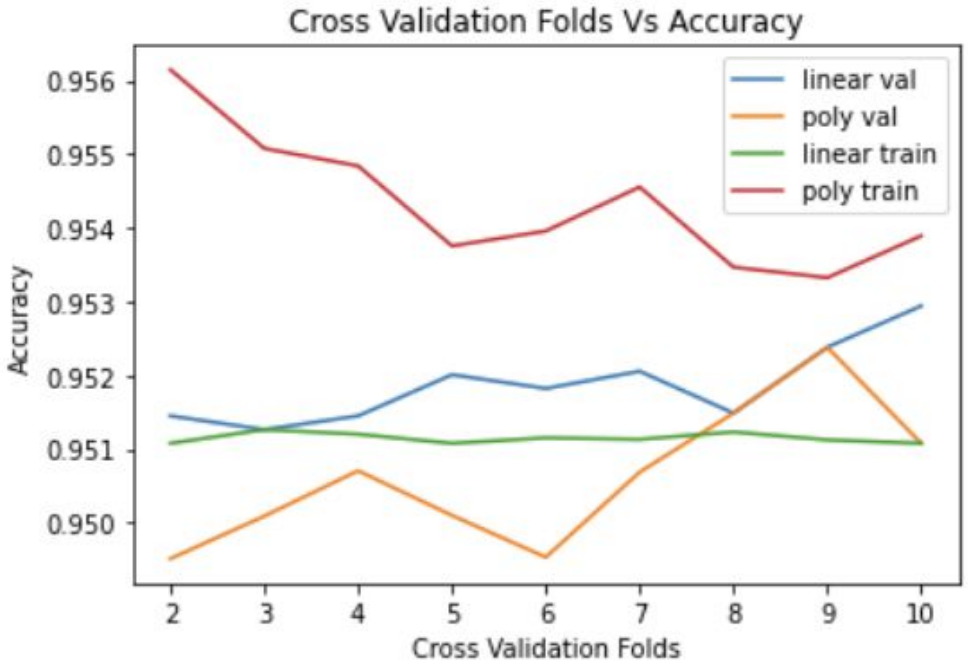
K-Fold Cross Validation

- 20% test set, 80% train and validation
- Folds: 2, 3, 4 … 10

Fold	poly val	poly train	linear val	linear train
2	0.949511	0.956147	0.951449	0.951076
3	0.950088	0.955079	0.951262	0.951262
4	0.950704	0.954842	0.951449	0.9512
5	0.950098	0.953756	0.952008	0.951076
6	0.949531	0.953958	0.951821	0.951151
7	0.950685	0.954556	0.952055	0.95113
8	0.951487	0.953468	0.951487	0.95123
9	0.952381	0.953325	0.952381	0.951123
10	0.951076	0.953893	0.952941	0.951076

K-Fold Cross Validation

- Best linear validation accuracy: 0.952941
- Best polynomial validation accuracy: 0.952381



Test Scores

- Slight improvement with cross validation
- Convergence around 0.94

Linear	Polynomial degree 1	Linear CV	Polynomial CV
0.941292	0.941292	0.941973	0.943901

Neural Network

1. No regularization on neural networks of 3, 4, and 5 layers

3 Layer: 17 - 10 - 1

4 Layer: 17 - 10 - 5 - 1

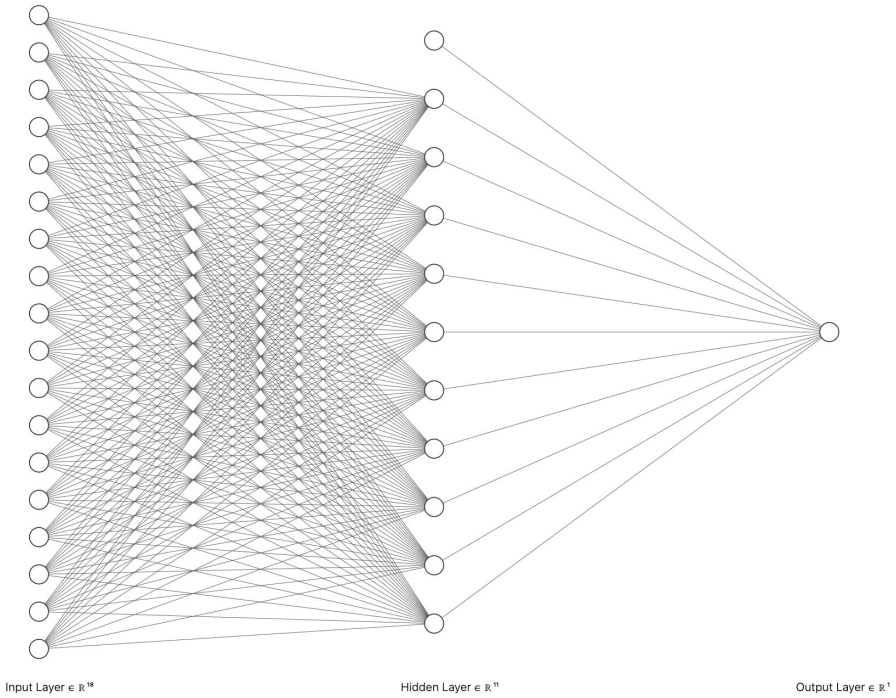
5 Layer: 17 - 12 - 7 - 4 - 1

2. L1 and L2 regularization on all neural networks above

Experiment Settings:

- Bias: Included
- Metric: Accuracy
- Batch size: 10, Epochs = 70

Three Layer Neural Network

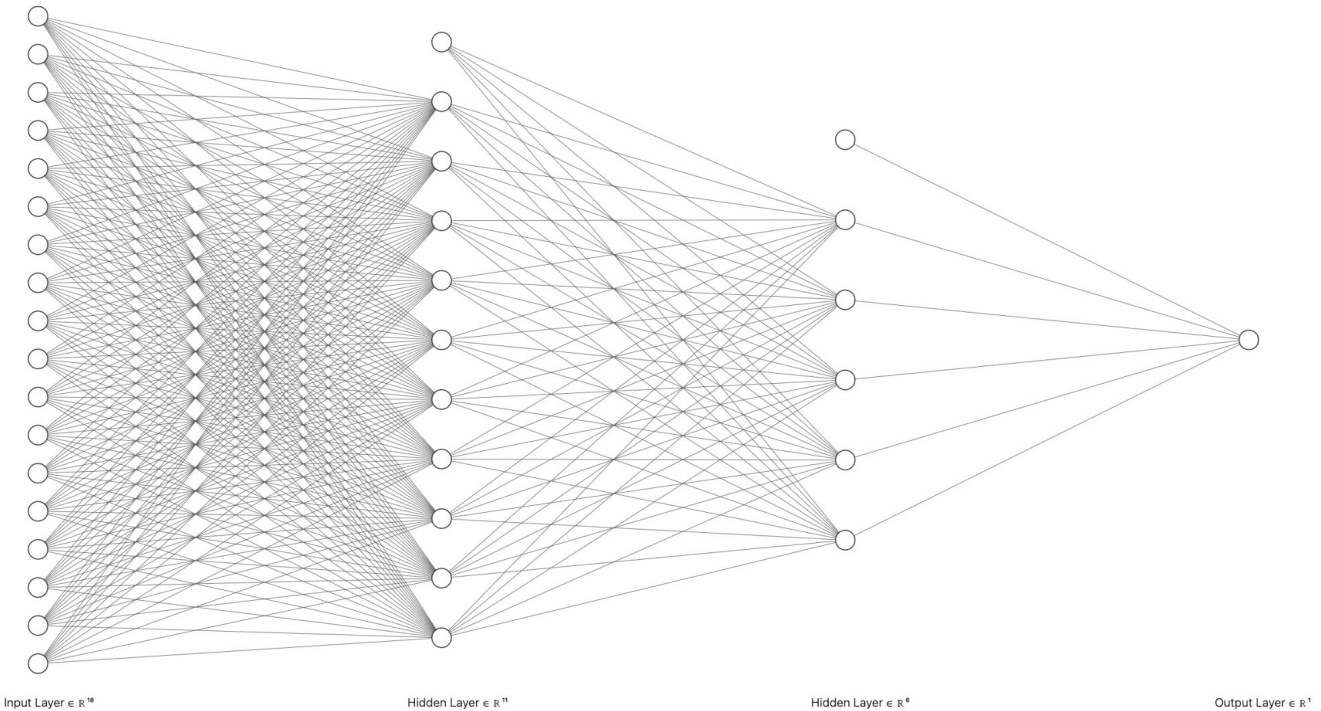


Three Layer Neural Network

3-layer Neural Network			
Metric	**Regularization**		
	No Regularization	***L1 Norm***	***L2 Norm***
Best Validation Loss	0.1719	0.1920	0.1877
Best Validation Accuracy	0.9499	0.9499	0.9499
Best Validation Epoch	24	69	68

Table 7: Three Layer Results

Four Layer Neural Network

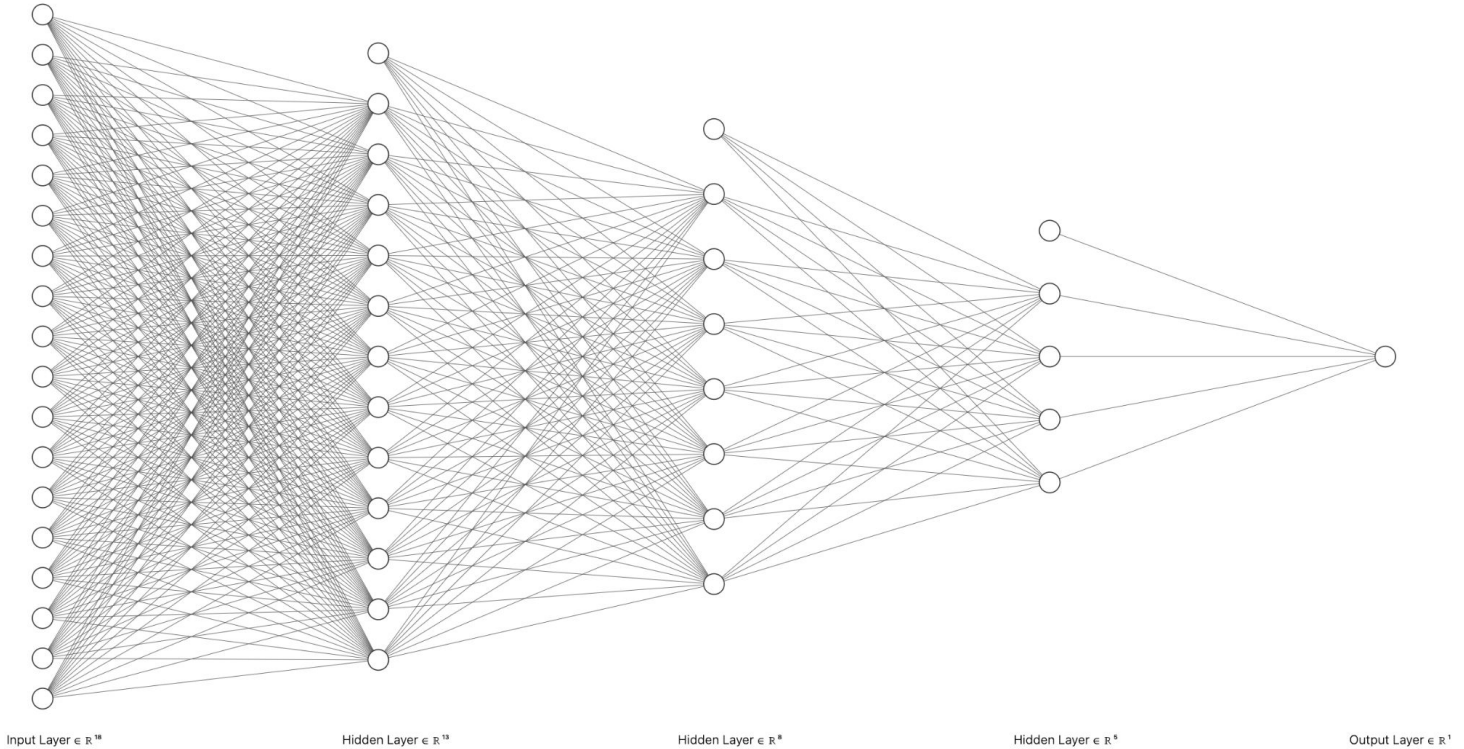


Four Layer Neural Network

4-layer Neural Network			
Metric	**Regularization**		
	No Regularization	***L1 Norm***	***L2 Norm***
Best Validation Loss	0.1687	0.2020	0.1967
Best Validation Accuracy	0.9499	0.9499	0.9499
Best Validation Epoch	9	69	51

Table 8: Four Layer Results

Five Layer Neural Network



Five Layer Neural Network

5-layer Neural Network			
Metric	**Regularization**		
	No Regularization	***L1 Norm***	***L2 Norm***
Best Validation Loss	0.1964	0.1964	0.1965
Best Validation Accuracy	0.9499	0.9499	0.9499
Best Validation Epoch	17	52	40

Table 9: Five Layer Results

Neural Network - Conclusions

Best Model: 4 layers, no regularization

Test loss and accuracy: 0.1808, 0.9413

Insights:

- Better with no regularization, but needs validation & early stopping
- Underfitting with regularization

Final Conclusion

- Polynomial kernel SVM with 9 fold cross validation wins!
- Dataset is extremely linearly separable,
- Potential Bottleneck:
  - Outliers in the dataset
  - Unaccounted/missing features in the dataset
  - Noise / variance in the problem itself

Logistic	SVM	Neural Network
0.9413	0.9439	0.9413

Logistic Regression & SVM Biggest Contributors

	0	1
13	Never_worked	-0.223147
4	ever_married	-0.164244
15	Self-employed	-0.138465
5	Residence_type	-0.078608
10	never smoked	-0.069239
12	Govt_job	-0.010489
11	smokes	0.007365
9	formerly smoked	0.020289
14	Private	0.040772
8	Unknown	0.050301
0	gender	0.055572
7	bmi	0.081896
3	heart_disease	0.087804
2	hypertension	0.089336
16	children	0.142904
6	avg_glucose_level	0.178882
1	age	1.660218