Dataset: Wisconsin Breast Cancer (Original)

Applying numerous machine learning algorythms

Structure of the presentation

- What I have done?
  - Logistic Regression
  - Neural Networks
  - Error, bias vs variance
- What does the future brings?
  - SVM
  - Diagnostic
  - Prognostic

First steps

- Number of instances: 699
- Number of attributes: 10 + class atribute
  - Sample code number
  - 9 attributes from 1 to 10 values.
  - Class attribute: 2 for benign, 4 for malignant
- Missing attributes: 16 instances
- Class distribution:
  - Benign: 458 (65.5%)
  - Malignant: 241 (34.5%)

First steps

- Number of instances: 699
- Number of attributes: 10 + class atribute
  - Sample code number
  - 9 attributes from 1 to 10 values.
  - Class attribute: 2 for benign, 4 for malignant
- Missing attributes: 16 instances
- Class distribution:
  - Benign: 458 (65.5%)
  - Malignant: 241 (34.5%)

How do we work with this dataset:

1. Reading the dataset file .data
2. Removing the "***Sample code number attribute***"
3. Transforming the output from 2 and 4 to 0 and 1

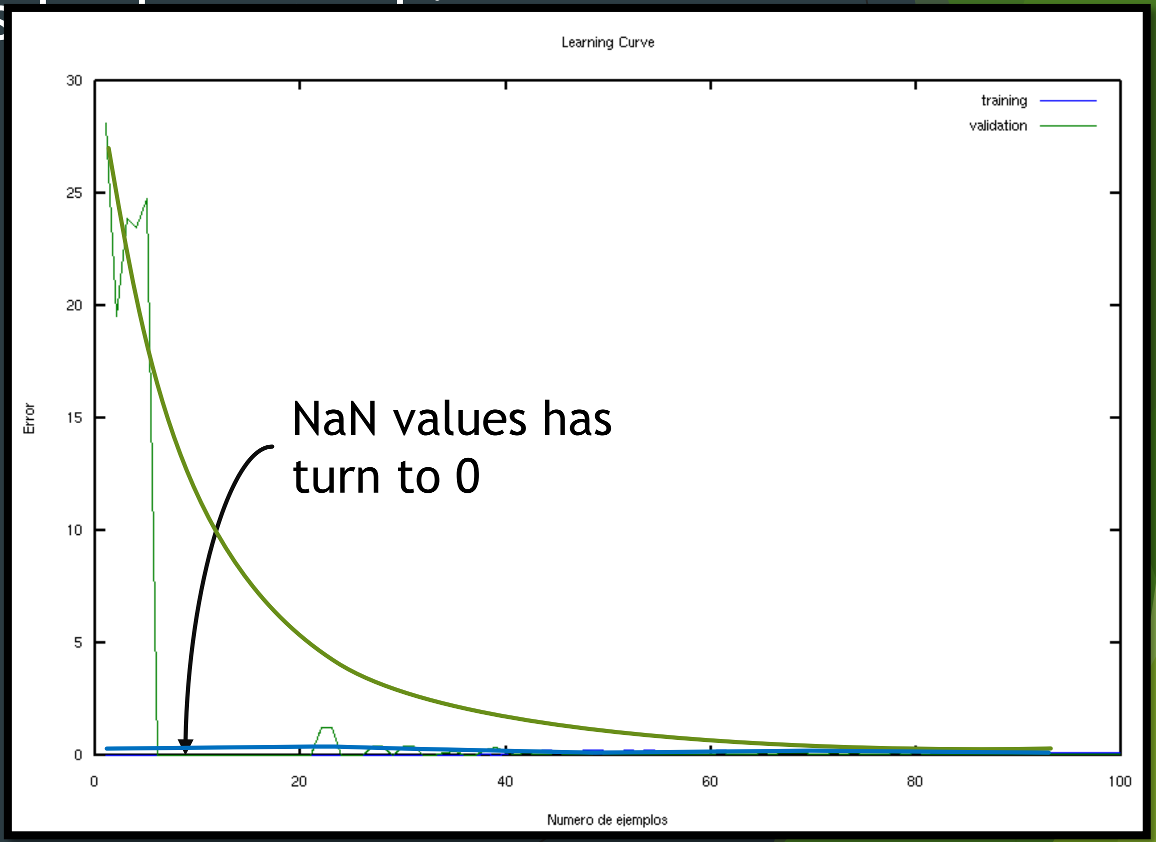
   Y = (Y == 4);

Logistic Regression

- With the data recently read, let’s try how good is our hypothesis.
- Without regularization.
- Splitting the dataset in two: 70% is for training data, 30% is for cross-validation data.

Logistic Regression

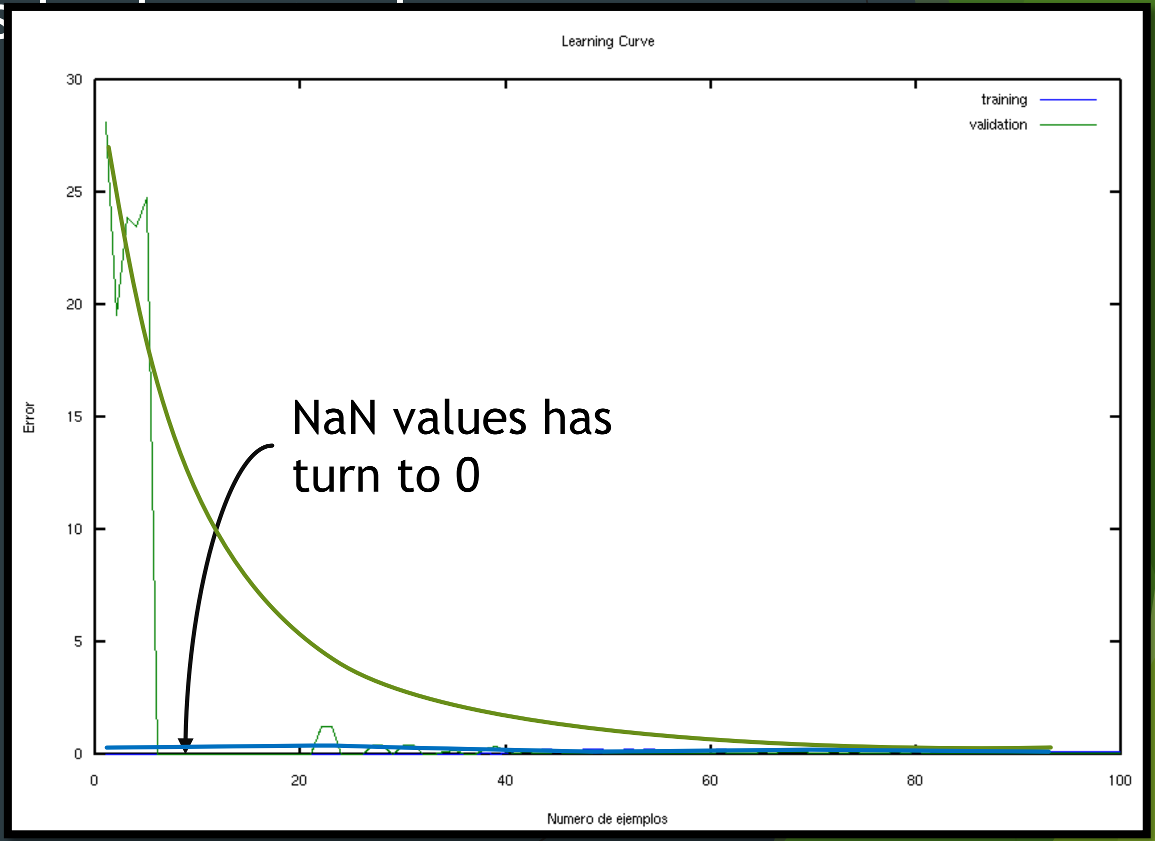
- With the data recently read, let's hypothesis.
- Without regularization.
- Splitting the dataset in two: 70% is for cross-validation data.



NaN values has turn to 0

Logistic Regression

- With the data recently read, let's
  hypothesis.
- Without regularization.
- Splitting the dataset in two: 70%
  is for cross-validation data.



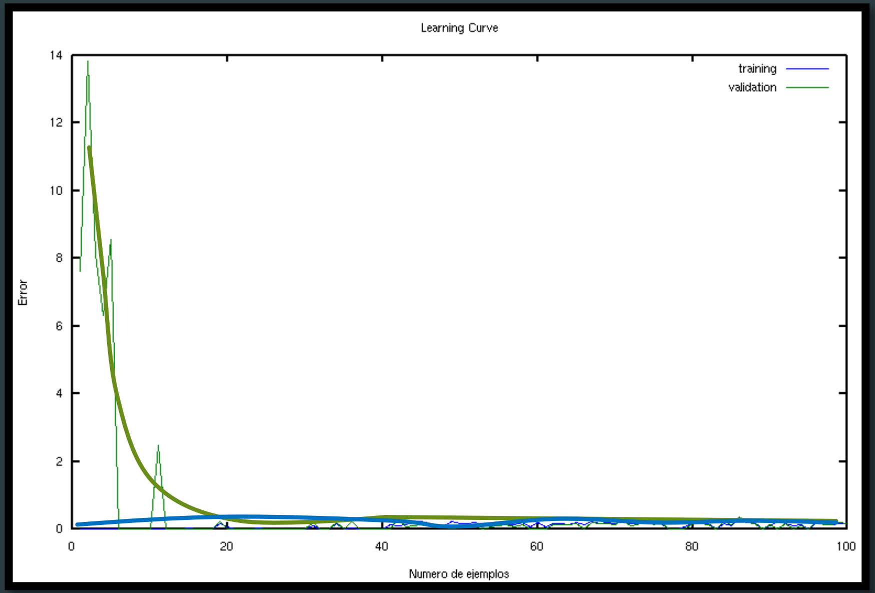
NaN values has turn to 0

We can do it better…

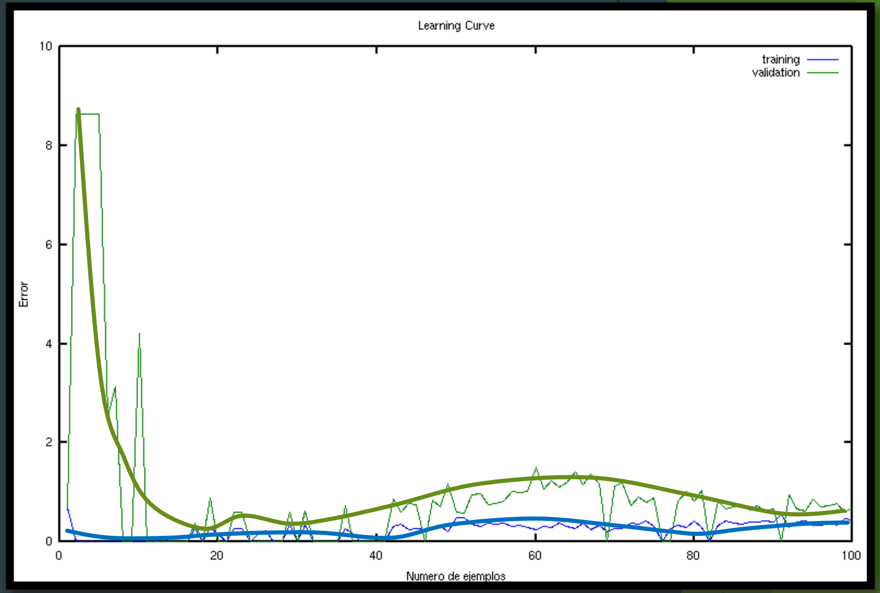
Logistic Regression

- Increasing the degree of our hypothesis (by increasing the number of features)
- Using a function that combine the features geometrically
  - X1,X2,X1*X2,X1^2,X2^2...

Logistic Regression



DEGREE=2



DEGREE=5

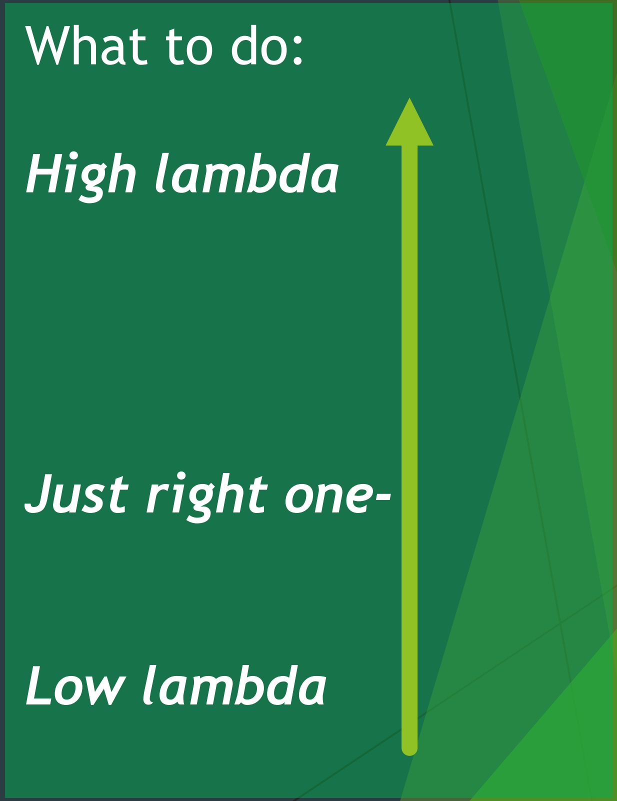
- You can check that an increasing in the degree of the polynomial implies a low bias between 40 and 80 dataset size, but high bias >80. (Increment of variance)

Logistic Regression

- We assume that:
  - $\lambda \uparrow$ fixes high bias ⟵ Our Problem
  - $\lambda \downarrow$ fixes high variance
- Now splitting in three the dataset:
  - 60% Training
  - 20% Cross-Validation
  - 20% Testing

Logistic Regression

- We assume that:
  - $\lambda \uparrow$ fixes high bias ⟵ Our Problem
  - $\lambda \downarrow$ fixes high variance
- Now splitting in three the dataset:
  - 60% Training
  - 20% Cross-Validation
  - 20% Testing



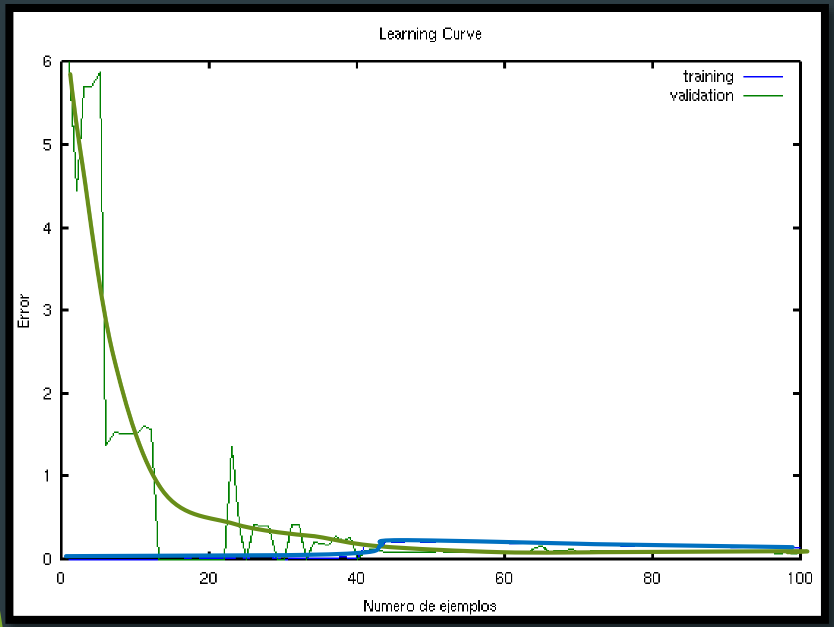
What to do:

High lambda

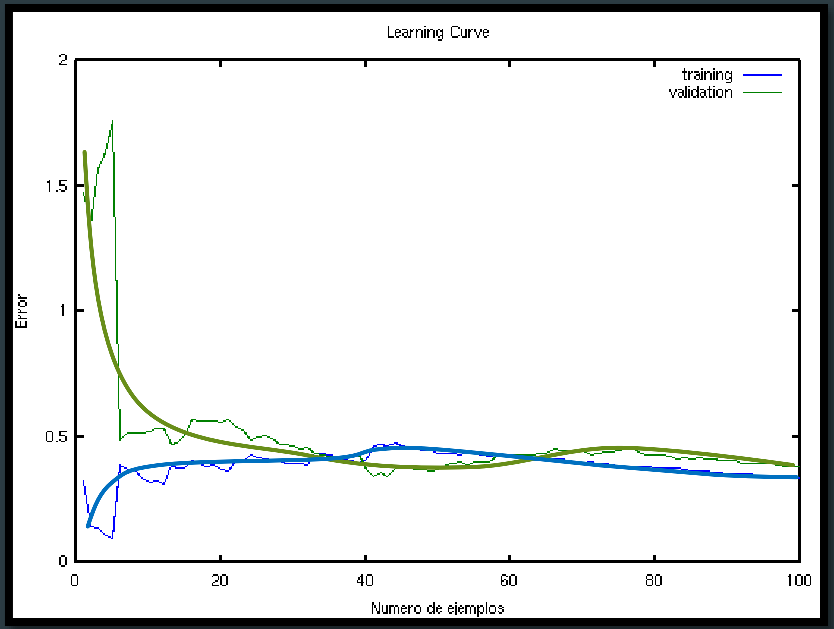
Just right one-

Low lambda

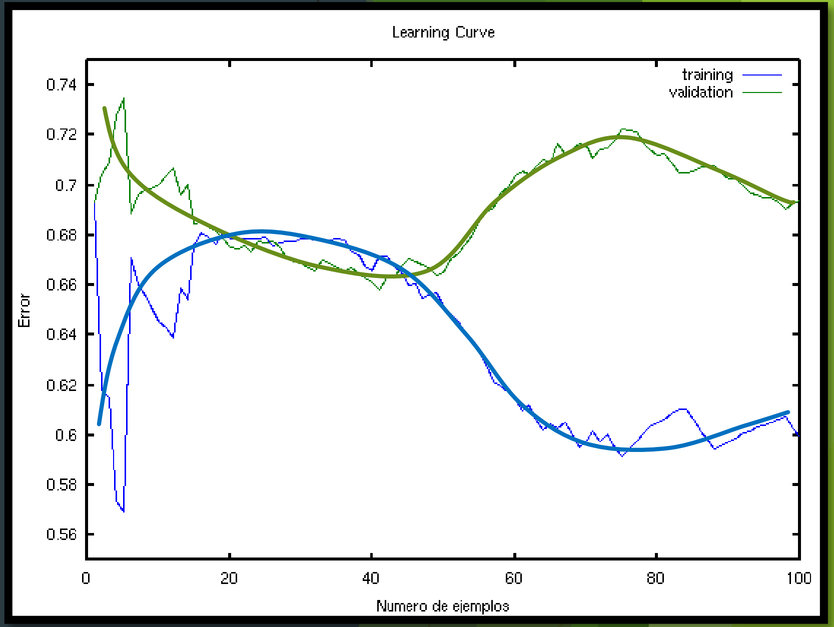
Logistic Regression



- Low lambda (0.001)
- High bias



- Just right lambda (3)
- The greatest of the 100% test result



- High lambda (300)
- High variance

Logistic Regression

- Just one thing left… Put it all together.
  - Looking for the best relation bias-variance.
  - Testing diferents splittings of the data to know what kind of split its the better.
  - Try to diagnose, by giving an hand-made example, if is begning or malignant.

Neural Networks

- Looking for the right structure of our network.
  - Best number of hidden units
  - Best number of hidden layers

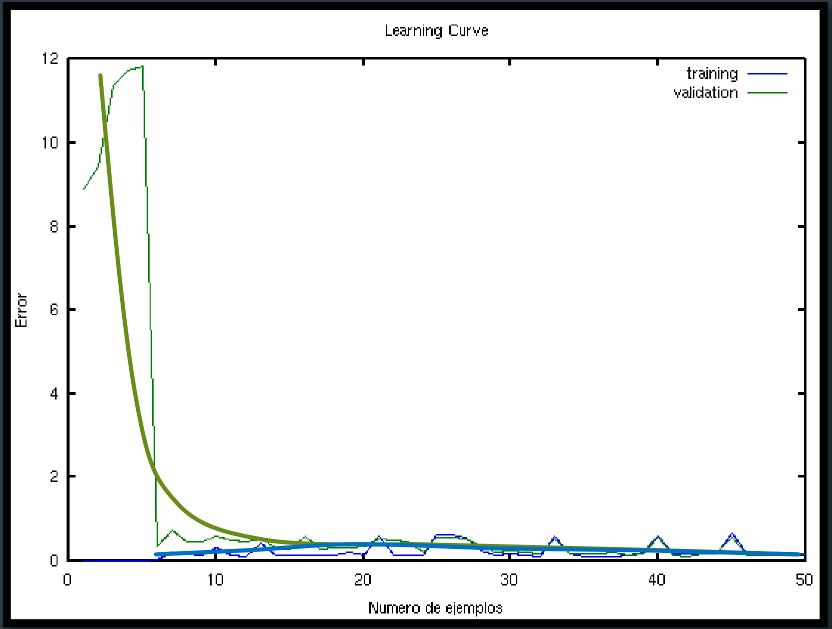
Neural Networks

- Looking for the right structure of our network.
  - Best number of hidden units
  - Best number of hidden layers

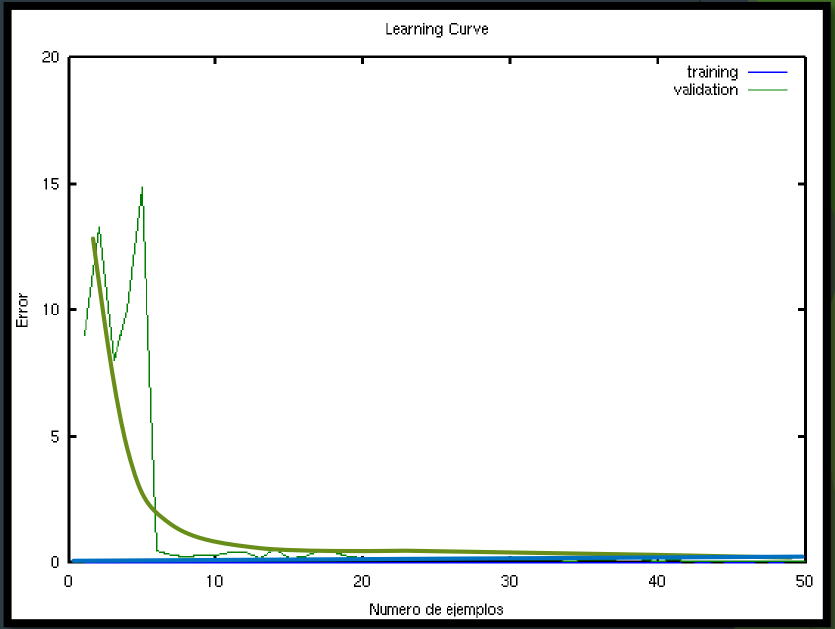
More hidden units/layers → High variance (overfitting)

Less hidden units/layers → High bias (underfitting)

Neural Network

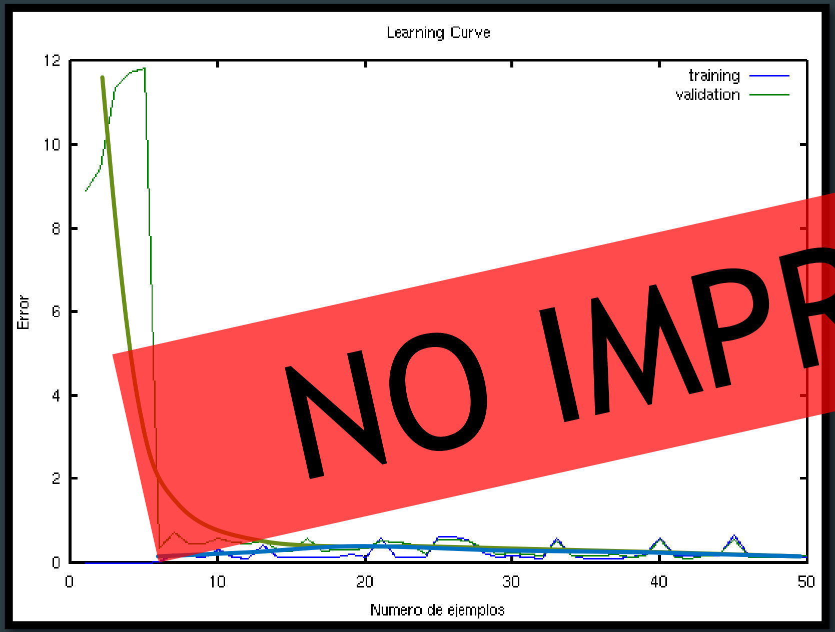


Lowest hidden units

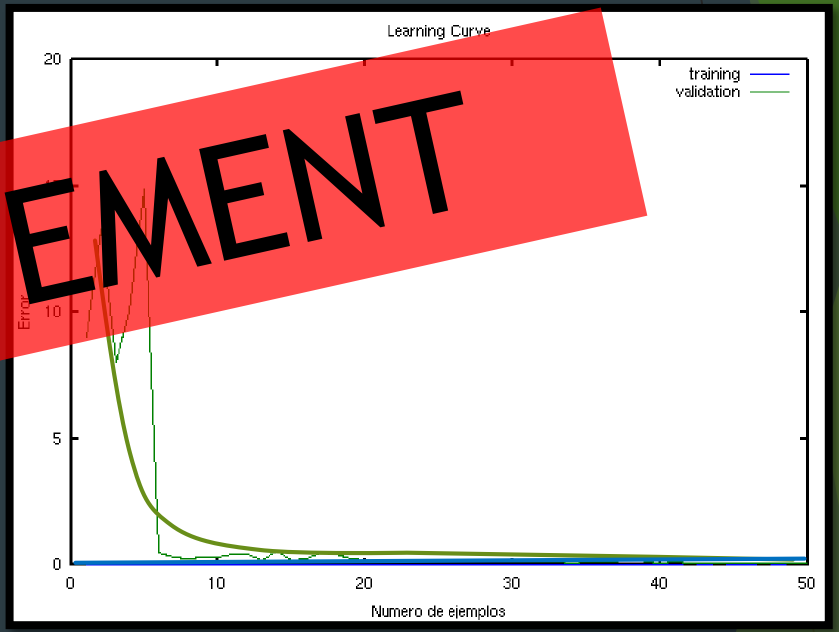


Highest hidden units

Neural Network



Lowest hidden units



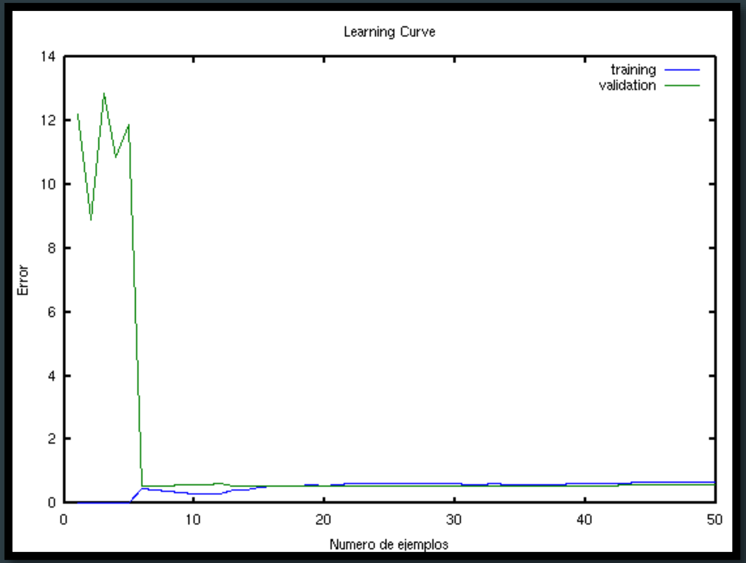
Highest hidden units

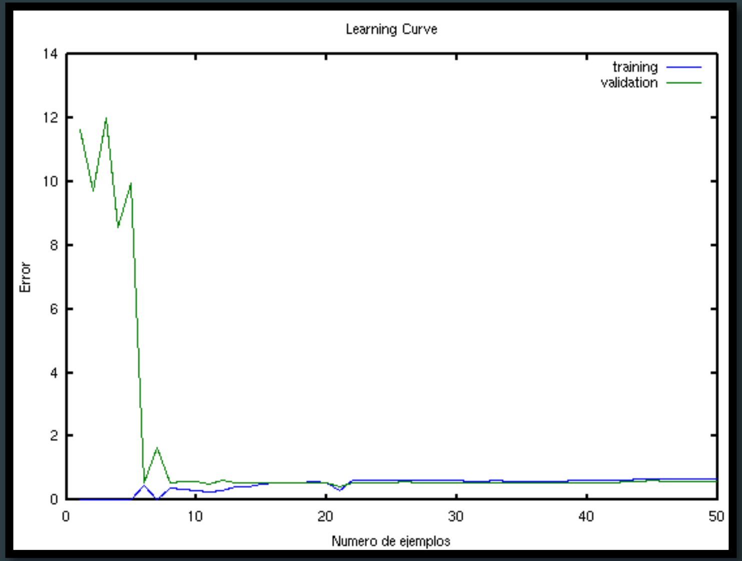
NO IMPROVEMENT

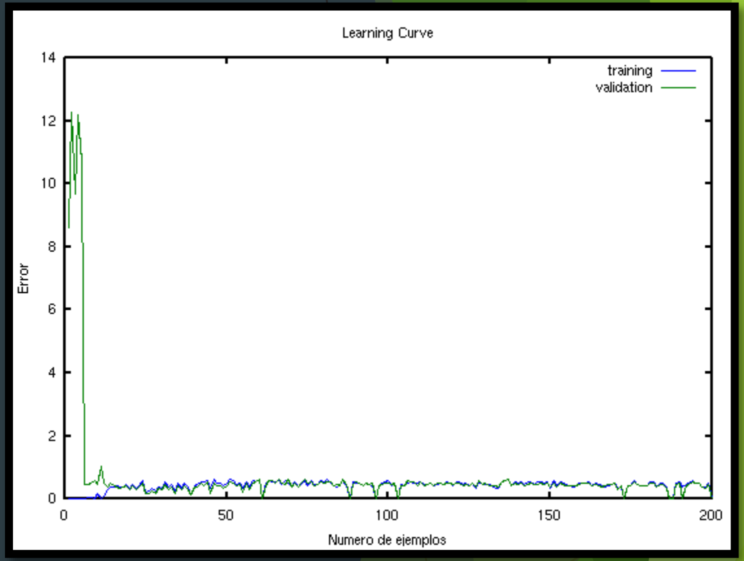
Neural Network

- Adding polynomial features reduces high bias.
- Pit it all together
  - Combine more hidden units/layers with more features.
  - Degrees 2, 3, 4 and 5.
  - Hidden Units 2, 4, 10, 50 and 100.

Neural Network



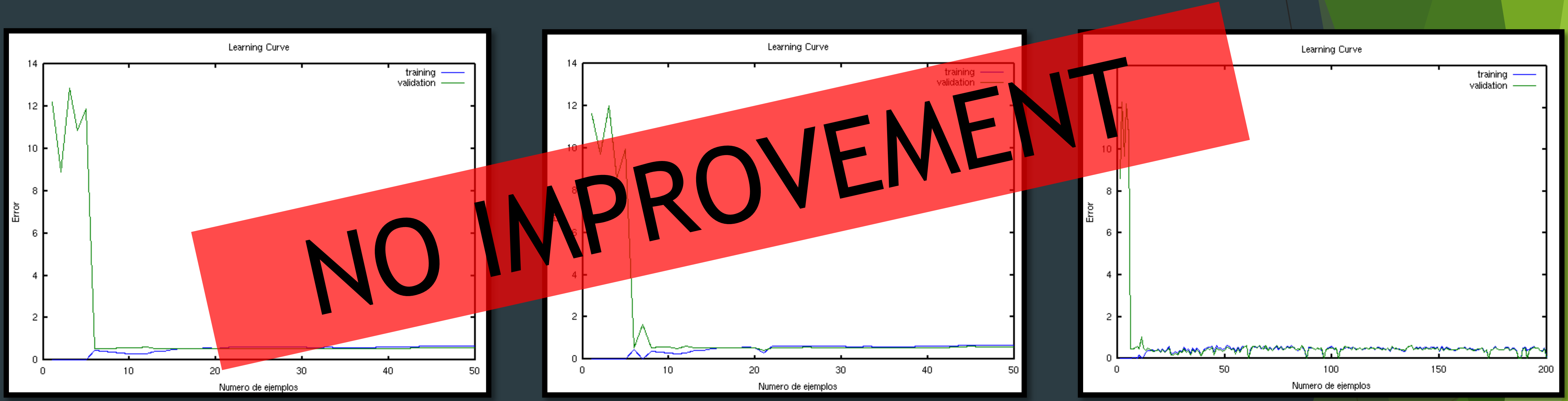




The same result as before.

More training sets

Neural Network



NO IMPROVEMENT

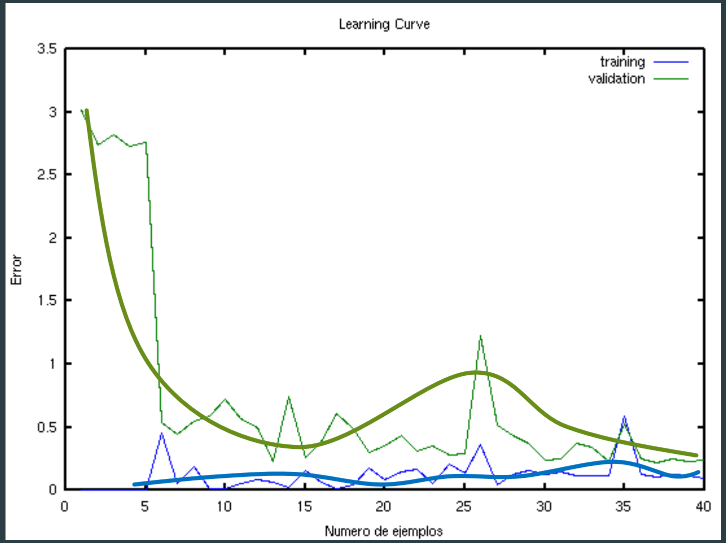
The same result as before.

More training sets

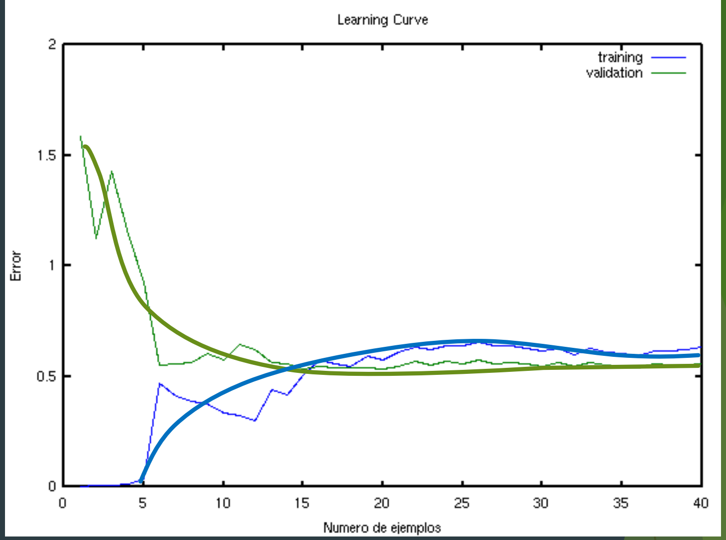
Neural Network

- Only thing left is to add the regularization parameter lambda.
- Remember:
  - $\lambda \uparrow$ fixes high bias ← Our Problem
  - $\lambda \downarrow$ fixes high variance

Neural Network



Lowest lambda



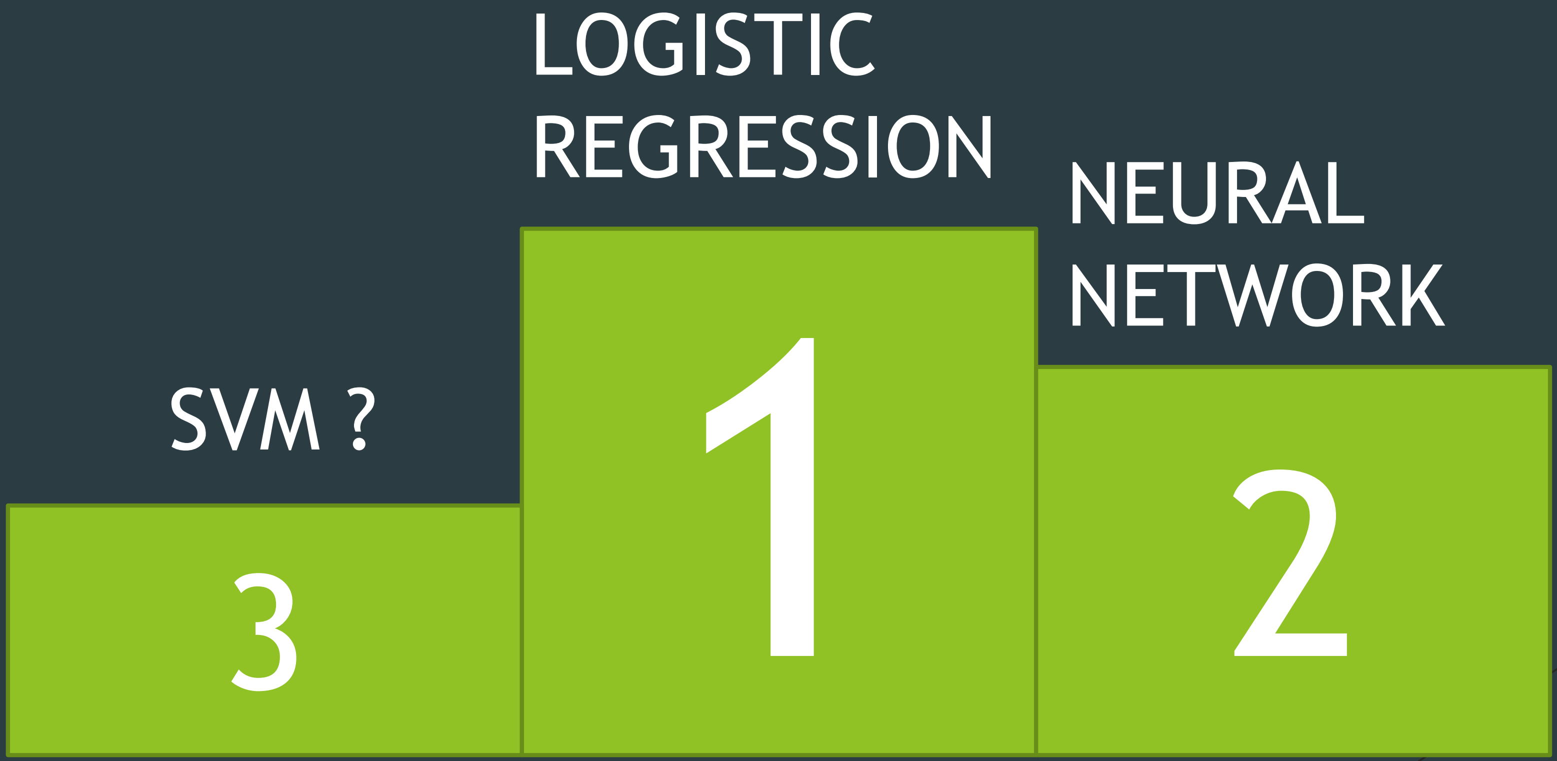
Highest lambda

A big difference between regularizated and non-regularizated.

Conclusions

- Regularization is far the most important thing to be aware of.
- Is positive to use polynomial adding (only if it worths).
- For every algorythm implemented, Error check is a must do.

Which is better?



SVM ?

3

LOGISTIC REGRESSION

1

NEURAL NETWORK

2

What does the future brings?

- SVM
  - Best C Parameter.
  - Which kernel is the better.
  - Check what percentage of data makes the best model.
- Diagnostic
  - Using the same algorythms as before.
- Prognostic
  - With all the data learning, to prognose if a patient would be regresive or no regresive.
- Maybe more…