

Applied Machine Learning, III

INFO 370

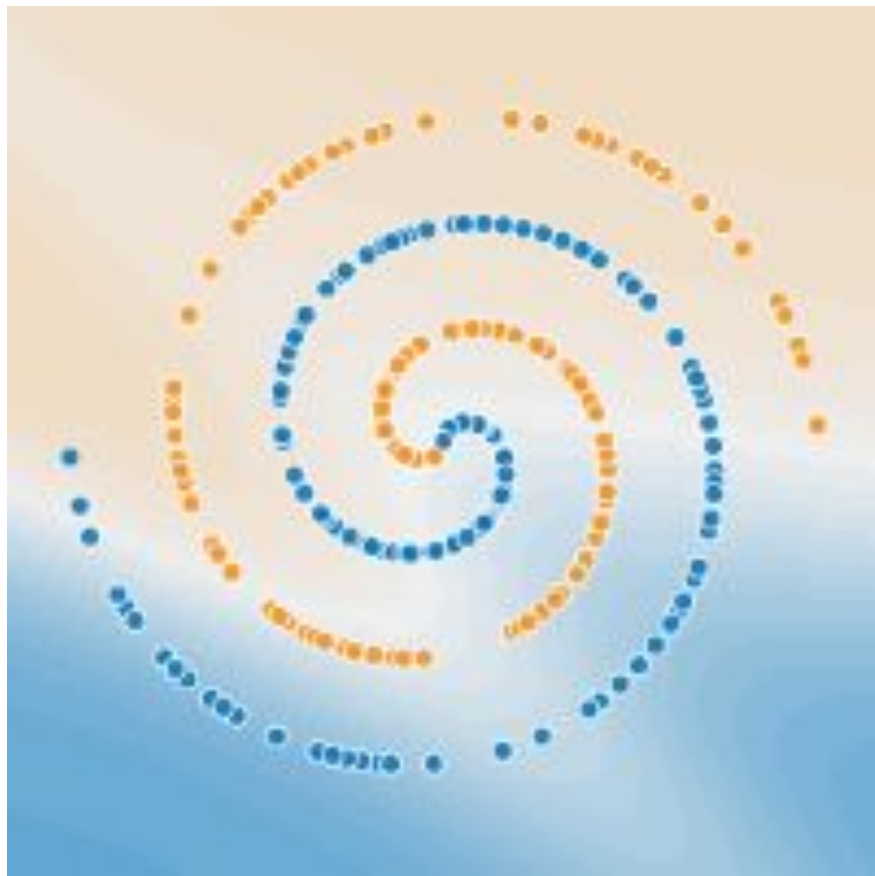
Learning Objectives

Discuss **neural networks**

Describe the importance **of feature selection**

Provide time to work on A4 // final project

Neural Networks

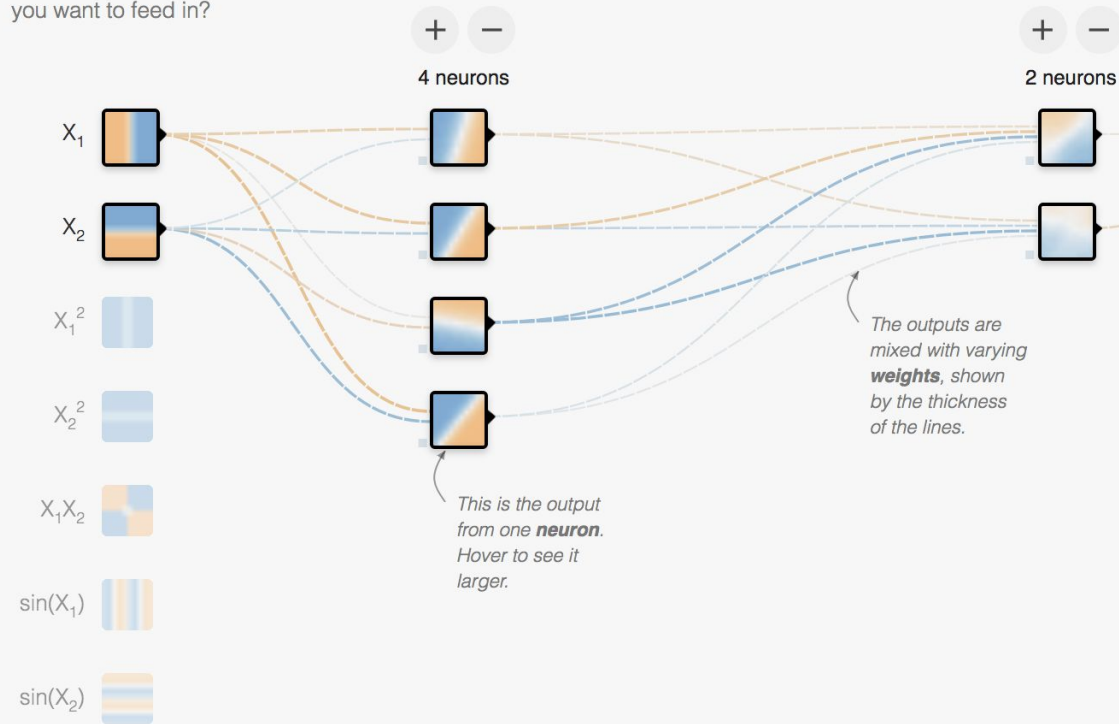


Why we need to talk about neural networks.

In Neural Networks,
Backpropagation performs a
gradient descent within the
solution's vector space
towards a 'global minimum'
along the steepest vector of
the error surface ([source](#)).

FEATURES

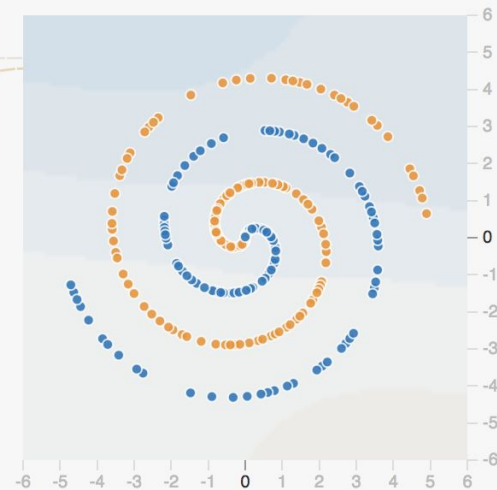
Which properties do you want to feed in?



OUTPUT

Test loss 0.513

Training loss 0.515



Makes more sense to explore ([link](#))

```
# Import and create a neural network classifier  
from sklearn.neural_network import MLPClassifier  
clf = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5, 2))
```

As you might imagine, sklearn implementation is trivial (but [tensorflow](#) is more popular)

Feature Selection

Why might it be
important to only
use certain (relevant)
features?

Feature Selection Strategies

Dimensionality reduction (PCA): transform current features into composite features

Feature selection: identify a heuristic for selecting a subset of "relevant" features

- Ignore features with low variation
- Select using Chi Squared Test
- Select using elements that emerge as relevant in a decision tree

How many features should you select?

- Use a GridSearch to figure it out!

```
# Using a variance threshold
from sklearn.feature_selection import VarianceThreshold
select_variance = VarianceThreshold(threshold=(.8 * (1 - .8)))
pipe = make_pipeline(imputer, scaler, select_variance, clf)

# Using SelectKBest -- or, more generally, SelectPercentile
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
select_k = SelectKBest(chi2)

# Figure out what K should be -- the number of features
param_grid['selectkbest__k'] = np.arange(10, 500, 30)
pipe = make_pipeline(imputer, scaler, select_variance, select_k, clf)

# Both!
pipe = make_pipeline(imputer, scaler, select_variance, select_features, clf)
```

Feature Selection Options

A4 Tips

Make sure you:

- Write **helper functions** to have a clean and **organized project**!
- Incorporate **individual data**
- Generate **dummy features** from categorical features
- **Impute** missing values
- Use a **Scaler** on your data
- Use **cross-validation** to search for optimal model parameters for you classifier
- Hint: use the **scorer** to match the competition scorer

The best strategy
will be to have a well
organized project.

Upcoming...

Assignment-4 due **Wednesday at 3:30**

Start chipping away at **your final projects (due in 1.5 weeks!)**