PURDUE UNIVERSITY | Department of Computer Science

CS57300: Data Mining

Neural Networks / Deep Learning

3 March 2022

Prof. Chris Clifton

PURDUE UNIVERSITY
Department of Computer Science

Hyperparameters

- Use as classifier: Thresholds
- Activation Function
- Gradient Descent
  - Starting weights
  - Learning rate
  - Epochs / stopping critera
- *Network structure*
  - May lead to additional parameters

PURDUE UNIVERSITY
Department of Computer Science

Plot Learning curve

- For a given dataset S, partition it into K folds S_1, S_2, …, S_K
- For frac = [10, 20, …, 100]
  For i = 1:K
  Test set = S_i
  Randomly sample frac% of S_{-i} to construct the training set S_{train}
  Learn model on S_{train} (as a reference, you can estimate the learned model's performance on
  S_{train}, record it as $perf\_t_{i, frac}$)
  Evaluate model's performance on S_i, record it as $perf\_v_{i, frac}$
- Plot the training set size vs. model performance
  - Given a specific frac, model's performance is captured by the mean and standard errors of [$perf\_v_{1, frac}$, $perf\_v_{2, frac}$, …, $perf\_v_{K, frac}$]

PURDUE UNIVERSITY
Department of Computer Science

Using Learning Curves

- Goal: Low error in practical use

PURDUE UNIVERSITY
Department of Computer Science

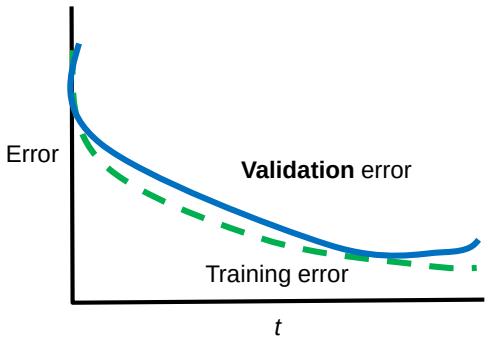
Epochs / Stopping Criteria

- Epoch: Pass through the training data
  - Batch GD: 1 step
  - Stochastic GD: Pass through the entire data (forward and backward)
- How many?
  - Fixed number?
  - Until error hits 0? Close to 0?
    - Training error?
    - Test error?

PURDUE UNIVERSITY
Department of Computer Science

Using Learning Curves

- In practice: Learning Curve vs. Epochs
  - Can we just stop at minimum test error?
- This overfits to test
  - Test error no longer a measure of expected accuracy
- *But still overfits to validation*



PURDUE UNIVERSITY
Department of Computer Science

Epoch / Stopping Critera

- Use minimum validation error
  - *In spite of overfitting*
- Stop slightly before minimum validation error
  - *How much before?*
- Use minimum validation error to determine number of epochs
  - "Final" training uses training and validation sets
  - *Nice if we are short on data, bad if short on computing power*

7

PURDUE UNIVERSITY
Department of Computer Science

Learning rate

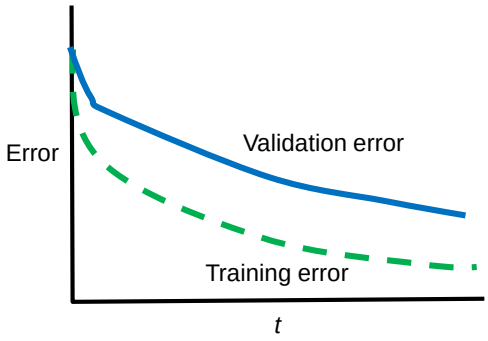
- High learning rate: Bigger steps
  - Gradient automatically reduces step size when we get close
  - So why not really large learning rate?
- Low learning rate: Small steps
  - Slower, but at least gives the right outcome
  - *Or does it?*
- Alternative: Reduce learning rate over time
  - *Ensure we at least fine* some *local minimum*

8

PURDUE UNIVERSITY
Department of Computer Science

Other Uses of Learning Curves

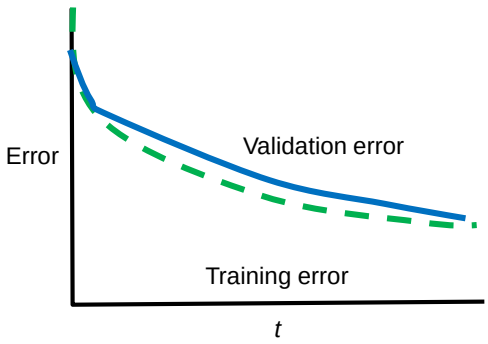
- Large gap: High Variance
  - Difference between error on different datasets
- Solutions
  - More data
  - Simpler model
  - Regularization



PURDUE UNIVERSITY
Department of Computer Science

Other Uses of Learning Curves

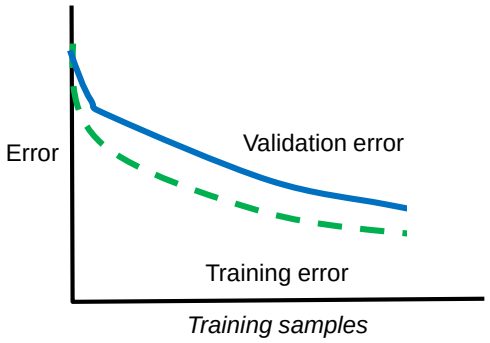
- Large error, small gap
  - High bias, low variance
- Solutions
  - More complex model



PURDUE UNIVERSITY
Department of Computer Science

Other Uses of Learning Curves

- Will more training data help?
  - Are errors decreasing with more training data?



PURDUE UNIVERSITY
Department of Computer Science

Initial Weights

- Why not 0?
  - $w_{kj} = -\sum_{i \in downstream(j)} (y_i^{(d)} - o_i^{(d)}) o_i^{(d)} (1 - o_i^{(d)}) w_{ji} o_j^{(d)} (1 - o_j^{(d)}) x_k^{(d)}$
- What about a constant?
  - *You'll think about this on the assignment*
- Typically: *small* random (close to 0, but non-zero)
  - Why random, when this prevents repeatability?

PURDUE UNIVERSITY
Department of Computer Science

Activation Function

- Sigmoid: Typically range (0,1) or (-1,1)
  - Monotonic, bell-shaped first derivative (steepest at input 0)
  - Logistic: $S(x) = \frac{1}{1+e^{-x}}$
  - Hyperbolic Tangent: $S(x) = \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$
  - Arctangent: $S(x) = \arctan x$
- Others possible – desirable properties?

PURDUE UNIVERSITY
Department of Computer Science

Threshold

- Not needed for regression
  - *We're trying to learn to predict a continuous value*
- But (differentiable) activation function gives a continuous value
  - What is the threshold for different classes?
- Don't we just learn this?
  - Set in the middle of the activation function...
- *But sometimes we may find adjusting helps*
  - What if Type I vs. Type II errors have different "value"

PURDUE UNIVERSITY
Department of Computer Science

ROC curves

- Receiver Operating Characteristic (ROC) curve
- Plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for different classification thresholds

P(Y)	True class
0.94	+
0.84	-
0.67	+
0.58	+
0.51	+
0.42	+
0.16	-
0.1	+
0.07	-

P(Y)	True class	Predict class
0.94	+	+
0.84	-	-
0.67	+	-
0.58	-	-
0.51	+	-
0.42	+	-
0.16	-	-
0.1	-	-
0.07	-	-

TPR = 1/4
FPR = 0/5

P(Y)	True class	Predict class
0.94	+	+
0.84	-	+
0.67	+	-
0.58	-	-
0.51	+	-
0.42	+	-
0.16	-	-
0.1	-	-
0.07	-	-

TPR = 1/4
FPR = 1/5

P(Y)	True class	Predict class
0.94	+	+
0.84	-	+
0.67	+	+
0.58	-	-
0.51	+	-
0.42	+	-
0.16	-	-
0.1	-	-
0.07	-	-

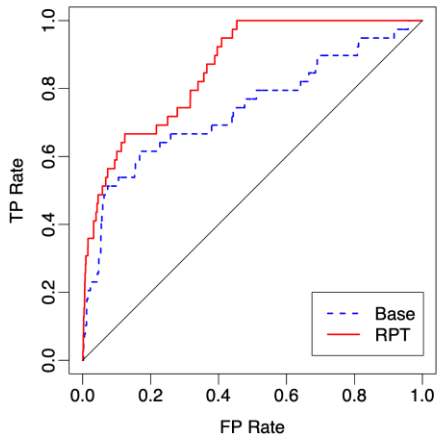
TPR = 2/4
FPR = 1/5

15

PURDUE UNIVERSITY
Department of Computer Science

AUC

- Evaluates performance over varying costs and class distributions
  - Can summarize with area under the curve (AUC)
  - AUC of 0.5 is random
  - AUC of 1.0 is perfect



16

PURDUE UNIVERSITY
Department of Computer Science

And then there is network structure…

- Number of input nodes, output nodes fixed by problem
- But number of hidden nodes can vary
  - Increase arbitrarily
- Can also increase number of hidden layers
  - Or have different inputs in different places
- And then there are *Recurrent* Neural Networks
  - E.g., LSTM

17

PURDUE UNIVERSITY
Department of Computer Science

Recurrent Neural Networks

- Used for processing *sequences* of inputs
  - Output is based on current "state" of network and inputs
- Example: LSTM
  - Long short-term memory
- Challenge: Training
  - Back-propagation can do weird things with "feedback" gradients

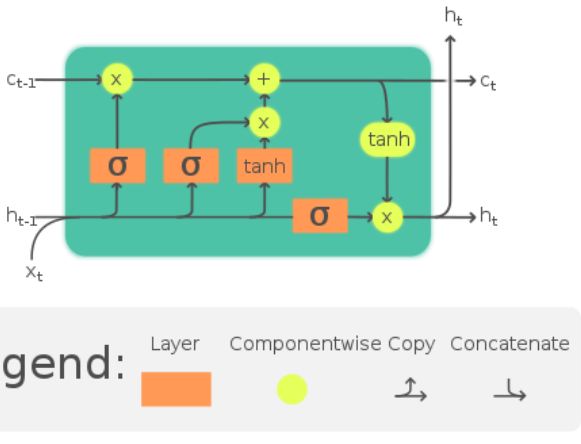


Image courtesy Guillaume Chevalier

18