Training Machine Learning Classifiers: Recap

Machine learning.

$$\Delta w_{ij} = -\alpha \frac{\partial E}{\partial w_{ij}}$$
Training/Validation/Test split

• Split the data into
  • Training set
    • Fit the classifier on the training data
  • Validation set
    • A “mock” test set: train different models, and run them on the validation set; pick the model that works best
      • “Model” can mean neural network architecture, or the parameters of the optimization, or the regularization parameters
  • Test set
    • Data that is held out and not used until the design process is over. Use for evaluating how the model will do on new data.
THIS IS JUST TO SAY

I have trained on
the data
that was in
the test set

and which
you were probably
saving
for validation

Forgive me
It reduced my
MSE
to nearly zero

#datascience #machinelearning #epitwitter
#statstwitter

7:29 PM · Jun 18, 2019 · Twitter Web Client
Training process

• For neural networks/logistic regression/linear regression, we train with gradient descent
• Obtain the training and validation cost at every iteration
  • Can also obtain the error (e.g. incorrect classification rate) at every iteration
Learning curves
Learning curves

\[ \varepsilon \]

Test error minimized

epochs

train

test
Stochastic gradient descent

- At every iteration, minimize the cost for a batch of data from the training set (rather than the entire training set)
- Easier computationally
- Usually works better
- "Stochastic" because at every iteration, there is a randomness element
  - We are not necessarily decreasing the training cost this way
    - Why?
"Oh sure, going in that direction will totally minimize the objective function" — Sarcastic Gradient Descent.
Regularization

• Want to do well on new data rather than on the training set
• There is sometimes a tradeoff
• Want to constrain the capacity of the classifier
  • It won’t do as well on the training set, but may do well on new data
• Methods
  • Early stopping: take the weights that minimize the cost on the validation set
  • L2 and L1 regularization: minimize cost+lambda*penalty
  • Train and average multiple models
  • Dropout
  • ...
• Usually want to regularize in some way
Belkin et al. (2019)