Minibatch Stochastic Gradient Descent
SpeakUp

Join by URL: https://web.speakup.info/room/join/80740

Join by key: web.speakup.info → use key 80740

Join by saved session: web.speakup.info

Join by QR:
Outline

• General neural network process

• (Batch) gradient descent

• Stochastic gradient descent (SGD)

• Minibatch stochastic gradient descent (Minibatch SGD)

• Discuss tradeoffs
Recap: Automatic Differentiation

• Take five minutes to draw
  • Whatever will help you remember (no correct or incorrect drawings)
  • You’ll keep a running drawing log the rest of the semester
\[
\text{mse: } (\hat{y} - y)^2 \rightarrow (\text{yhat} - y)^{**2}, \text{mean}()
\]

\[
\text{mae: } |\hat{y} - y| \rightarrow (\text{yhat} - y), \text{abs()}, \text{mean}()
\]

```python
matrix.py

```python
def abs(self) -> Matrix:
    result = ...

def -grad():
    result, grad = grad
    return result
```

```python
list2d.py

```python
def abs(self) -> list2d:
    for for for for return ...
```
\[ z = x + c \]
General Neural Network Process

1. Prepare your dataset
   • Normalize inputs
   • Create a proxy (smaller) dataset for debugging and testing workflows
   • Split the dataset into training, validation, and evaluation

2. Design a neural network (architecture)

3. Set initial hyperparameters (learning rate, number of epochs, etc.)

4. Train model parameters (includes validation)

5. Evaluate the trained model

6. Deploy the trained model

Software Engineering for Machine Learning: A Case Study

- Salimah Anani
  Microsoft Research
  Redmond, WA, USA
  salimah@msn.com
- Andrew Ng
  Microsoft Research
  Redmond, WA, USA
  andrewng@msn.com
- Christian Rohr
  Microsoft Research
  Redmond, WA, USA
  christianr@msn.com
- Robert DeLine
  Microsoft Research
  Redmond, WA, USA
  rdeline@msn.com
- Harold Gall
  University of Zurich
  Zurich, Switzerland
  gall@inf.tau.ch
- Farzad Kamnir
  Microsoft Research
  Redmond, WA, USA
  farzadm@msn.com
- Nathaniel Nguyen
  Microsoft Research
  Redmond, WA, USA
  niny@msn.com
- Shenzi Zhou
  Microsoft Research
  Redmond, WA, USA
  shenzizh@msn.com
- Thomas Zimmermann
  Microsoft Research
  Redmond, WA, USA
  thomaszm@msn.com
- Qiaochu Yang
  Microsoft Research
  Redmond, WA, USA
  qiaochu@msn.com
(Batch) Gradient Descent

for each epoch, we update the parameters while considering every training input example.

1. compute gradients w.r.t. all examples
2. average gradients for each parameter individually
3. update each parameter using its average gradient
4. validate results (check for overfitting)
for each epoch
  shuffle data
  for each example
    1. compute gradients
    2. update parameters
    3. validate
Minibatch Stochastic Gradient Descent

for each epoch
  Shuffle
  for each batch
    1. compute gradients
    2. average gradients
    3. update parameters
  4. validate

batch size
Typical Behavior

![Graph showing typical behavior with loss on the y-axis and epoch on the x-axis. The graph shows a general downward trend with fluctuations.](image-url)
Tradeoffs

• (Batch) gradient descent
  • Fewer updates means slower progress in terms of real time
  • Maximally leverages parallel computations (faster)

• Stochastic gradient descent
  • Many updates can sometimes lead to faster convergence
  • Noisy data means that some updates are bad

• Minibatch stochastic gradient descent
  • Noisy but less than SGD
  • Pretty fast but slower than batch gradient descent
  • Should always be your default choice
  • Add another hyperparameter called `batch_size`
Code Demo

https://github.com/anthonyjclark/cs152sp23/blob/main/Lectures/06-MinibatchSGDDemo.ipynb