CS540 Introduction to Artificial Intelligence
Lecture 4

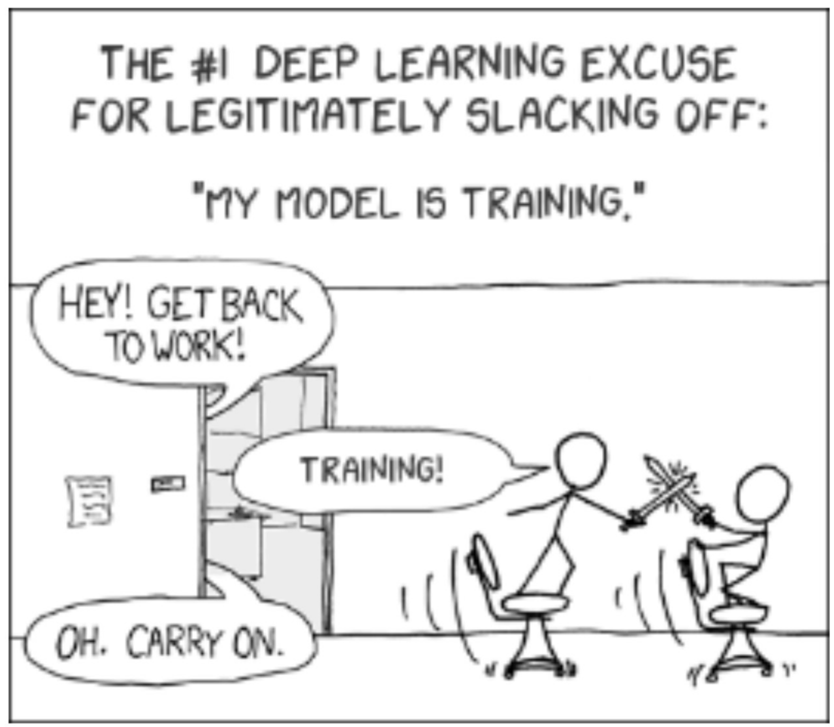
Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 5, 2020

Socrative Test

Admin

- Socrative Student Login: Room CS540C. Use the wisc.edu ID without the wisc.edu.
- Use Socrative Room CS540 (without the C) for anonymous feedback.
- A: I haven't started P1.
- B: I have started P1.
- C: I have finished part 1.
- D: I have finished P1.
- E: What is P1?



Perceptron Algorithm vs Logistic Regression

Motivation

- For LTU Perceptrons, w is updated for each instance x_i sequentially.

$$w = w - \alpha \left(a_i - y_i \right) x_i$$

- For Logistic Perceptrons, w is updated using the gradient that involves all instances in the training data.

NN $\Longrightarrow$

$$w = w - \alpha \sum_{i=1}^{n} \left(a_i - y_i \right) x_i$$

Stochastic Gradient Descent Diagram 1

Motivation

$$w = w - \alpha \sum_{i=1}^{n} \frac{\partial C_i}{\partial w}$$

samples

L3

$$\sum \underbrace{(\hat{y}_i - y_i)^2}_{C_i}$$

P1 logistic

batch GD

w

n

$\sum$

w'

all $\frac{\partial C_i}{\partial w}$

one

based on w

@ nth

P1 nn

SGD

w

$\frac{\partial C_1}{\partial w}$

w'

w''

$\frac{\partial C_2}{\partial w}$

based on w

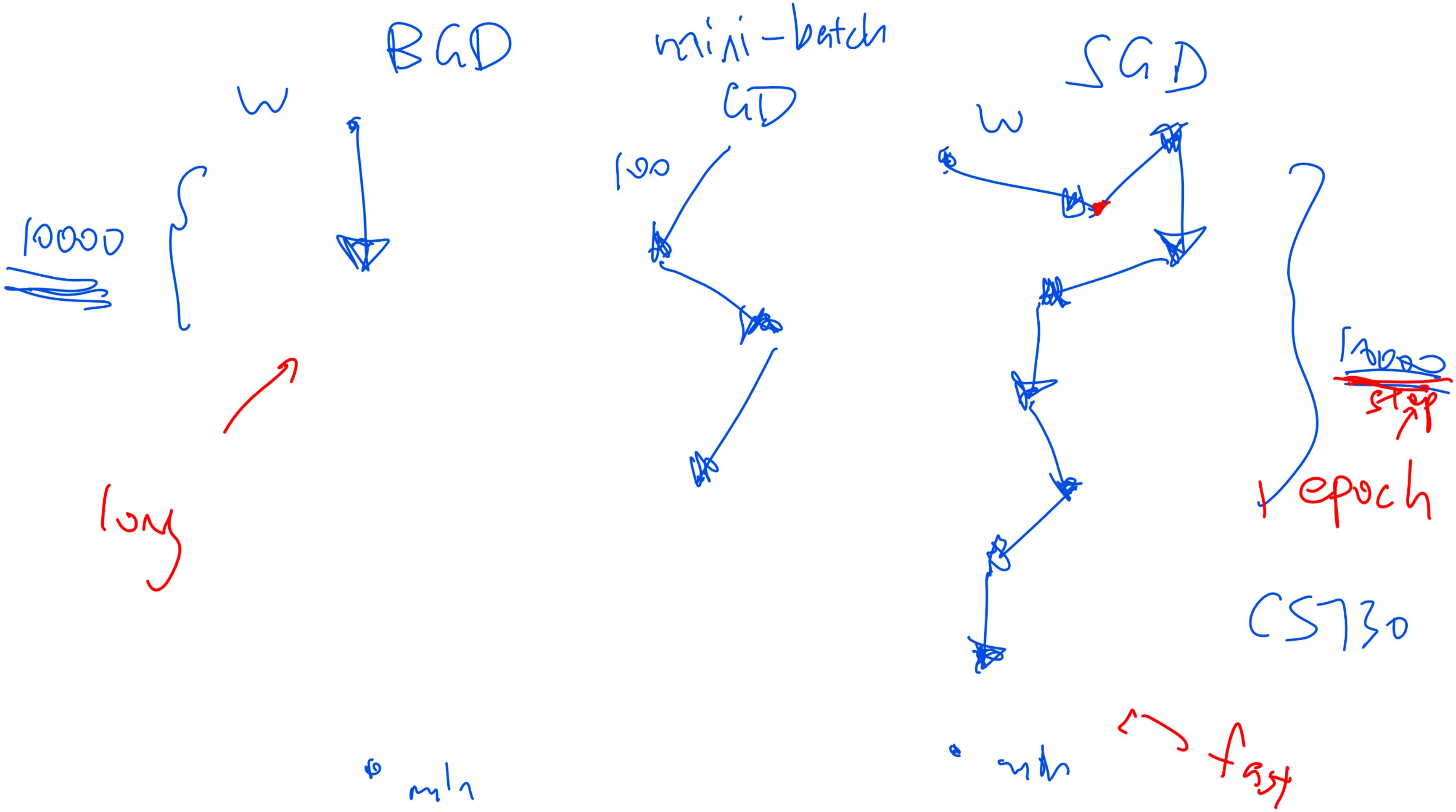
based on w'

n times

@ nth

Stochastic Gradient Descent Diagram 2

Motivation



Multi-Class Classification

Motivation

- When there are K categories to classify, the labels can take K different values, $y_i \in \{1, 2, ..., K\}$.
- Logistic regression and neural network cannot be directly applied to these problems.

Method 1, One VS All

Discussion

0.7

0 1 2

0.8

- Train a binary classification model with labels $y_i' = \mathbb{1}_{\{y_i = j\}}$ for each $j = 1, 2, ..., K$.
- Given a new test instance x_i, evaluate the activation function $a_i^{(j)}$ from model j.

$$\hat{y}_i = \arg\max_j a_i^{(j)}$$

- One problem is that the scale of $a_i^{(j)}$ may be different for different j.

Method 2, One VS One

Discussion

0 1 2

- Train a binary classification model with for each of the $\frac{K(K-1)}{2}$ pairs of labels.
- Given a new test instance x_i, apply all $\frac{K(K-1)}{2}$ models and output the class that receives the largest number of votes.

$$\hat{y}_i = \arg\max_j \sum_{j' \neq j} \hat{y}_i^{(j \text{ vs } j')}$$

$\boxed{\begin{matrix} 0 \\ 1 \\ 2 \end{matrix}}$

- One problem is that it is not clear what to do if multiple classes receive the same number of votes.

One Hot Encoding

Discussion

- If y is not binary, use one-hot encoding for y.
- For example, if y has three categories, then

$$y_i \in \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

Method 3, Softmax Function

Discussion

- For both logistic regression and neural network, the last layer will have K units, a_{ij}, for $j = 1, 2, ..., K$ and the softmax function is used instead of the sigmoid function.

$$a_{ij} = g\left(w_j^T x_i + b_j\right) = \frac{\exp\left(-w_j^T x_i - b_j\right)}{\sum_{j'=1}^{K} \exp\left(-w_{j'}^T x_i - b_{j'}\right)}, j = 1, 2, ..., K$$

Softmax Derivatives

Discussion

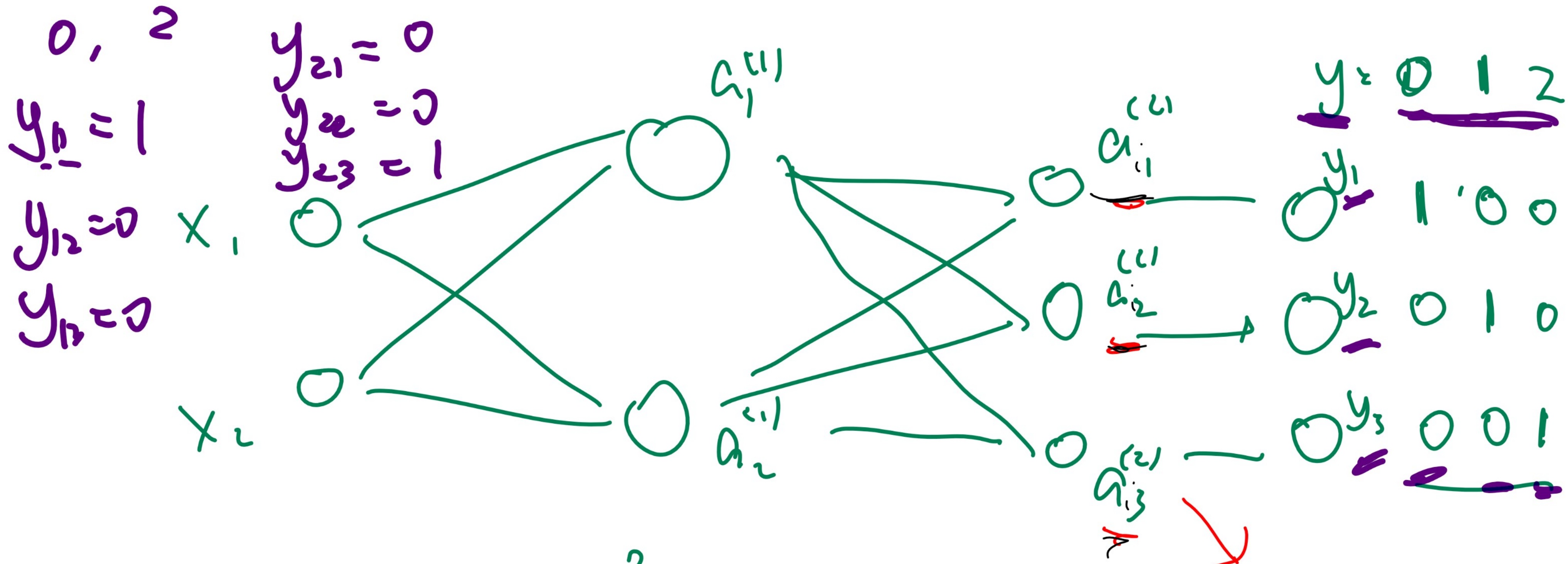
- Cross entropy loss is also commonly used with softmax activation function.
- The gradient of cross entropy loss with respect to a_{ij}, component j of the output layer activation for instance i has the same form as the one for logistic regression.

$$\frac{\partial C}{\partial a_{ij}} = a_{ij} - y_{ij} \Rightarrow \nabla_{a_i} C = a_i - y_i$$

- The gradient with respect to the weights can be found using the chain rule.

Softmax Diagram

Discussion



$$C = \sum_{i=1}^{n} \sum_{j=1}^{3} \frac{1}{2}\left(a_{ij}^{(2)} - y_{ij}\right)^2$$

$j \to 1, 2, 3$

$a_3^{(2)} = g(w^T a^{(1)} + b)$

~~logistic~~

softmax

$a_i - y_i \to \mathbb{R}^3$

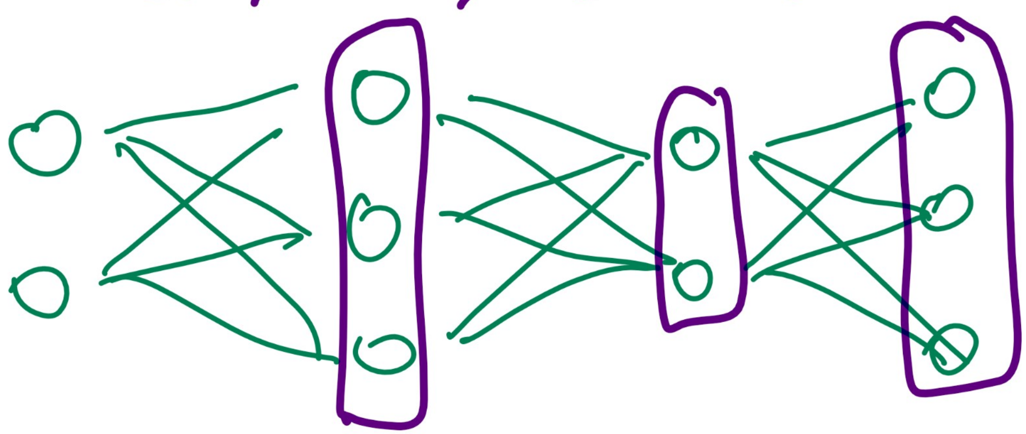
Weight Count

Quiz

← for each non-input unit

- How many weights and biases are there in a (fully connected) three layer neural network with 2 input units, 3 hidden units in the first hidden layer, 2 hidden units in the second hidden layer, and 3 output units?

2×3 3×2 2×3 18 weights



8 biases.

26 things to update.

Weight Count 2

Quiz

Q2

- How many weights (not including bias) are there in a (fully connected) two layer neural network with 10 input units, 5 hidden units, and 10 output units.
- A: 50
- B: 55
- C: 100
- D: 110
- E: 500

10 5 10

50 + 50 = 100

Weight Count 3

Quiz

Q3

- How many biases are there in a (fully connected) two layer neural network with 10 input units, 5 hidden units, and 10 output units.
- A: 5
- B: 10
- C: 15
- D: 20
- E: 25

10

$\boxed{5} + \boxed{10} = 15$

Questions about P1

Admin

$\frac{1}{2}(y_i - a_i)^2$

training { train, validate

$a = \sigma(w^T x + b)$

- Cost function? ✓
- Learning rate? $\alpha/\sqrt{t}$ ← t-th epoch
- Stopping criterion? $C < 0.01$ not recomended → $|C_t - C_{t-1}| < 0.001$ converge
- No Stochastic vs regular gradient descent?
- Regularization? ✓ λ small
- Use test set to train? NO.
- Other questions?

① $\frac{\partial C}{\partial w^{(2)}} =$

$\frac{\partial C}{\partial w^{(1)}} = (a_i - y_i)\, a_i^{(2)}(1 - a_i^{(2)})\, w^{(2)}\, a_i^{(1)}(1 - a_i^{(1)})\, x$

CS730 L3

② $w = w - \alpha \frac{\partial C}{\partial w}$

Generalization Error Diagram

Motivation

Method 1, Validation Set

Discussion

- Set aside a subset of the training set as the validation set.
- During training, the cost (or accuracy) on the training set will always be decreasing until it hits 0.
- Train the network until the cost (or accuracy) on the validation set begins to increase.

Method 2, Drop Out

Discussion

- At each hidden layer, a random set of units from that layer is set to 0.
- For example, each unit is retained with probability $p = 0.5$. During the test, the activations are reduced by $p = 0.5$ (or 50 percent).
- The intuition is that if a hidden unit works well with different combinations of other units, it does not rely on other units and it is likely to be individually useful.

Method 3, L1 and L2 Regularization

Discussion

- The idea is to include an additional cost for non-zero weights.
- The models are simpler if many weights are zero.
- For example, if logistic regression has only a few non-zero weights, it means only a few features are relevant, so only these features are used for prediction.

Method 3, L1 Regularization

Discussion

- For L1 regularization, add the 1-norm of the weights to the cost.

$$
\begin{aligned}
C &= \sum_{i=1}^{n} (a_i - y_i)^2 + \lambda \left\| \begin{bmatrix} w \\ b \end{bmatrix} \right\|_1 \\
&= \sum_{i=1}^{n} (a_i - y_i)^2 + \lambda \left(\sum_{i=1}^{m} |w_i| + |b| \right)
\end{aligned}
$$

logistic regression

force ws to be 0

- Linear regression with L1 regularization is called LASSO (least absolute shrinkage and selection operator).

feature selection

Method 3, L2 Regularization

Discussion

- For L2 regularization, add the 2-norm of the weights to the cost.

$$
\begin{aligned}
C &= \sum_{i=1}^{n} (a_i - y_i)^2 + \lambda \left\| \begin{bmatrix} w \\ b \end{bmatrix} \right\|_2^2 \\
&= \sum_{i=1}^{n} (a_i - y_i)^2 + \lambda \left(\sum_{i=1}^{m} w_i^2 + b^2 \right)
\end{aligned}
$$

Method 4, Data Augmentation

Discussion

- More training data can be created from the existing ones, for example, by translating or rotating the handwritten digits.