3. Training neural networks (Pan)

Covered topics

1. Gradient problems
2. Weight initialization
3. Batch normalization
4. L2 and L1 regularization
5. Dataset augmentation
6. Multitask and transfer learning
7. Ensemble methods
8. Dropout
9. Other techniques discussed in the lectures

Review materials

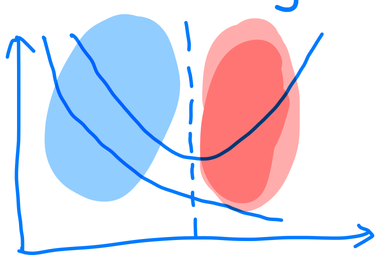
1. Lecture slides (formal notes)
2. Homework
3. Discussion handouts
4. Textbook (optional)

Problem examples

1. Which of the following techniques can be used to reduce overfitting? **Select all that apply**.

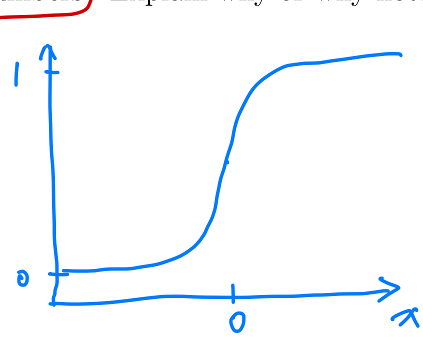
✓ A. Batch normalization
✓ B. Data augmentation
✓ C. Dropout
✗ D. Use less training data
✗ E. None of the above

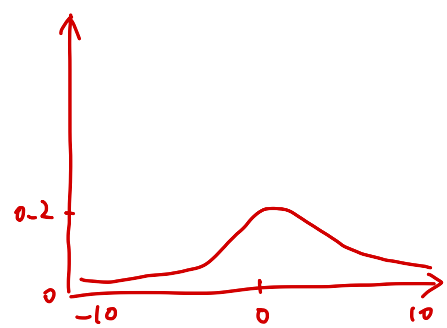
overfitting : good performance on the training data
bad performance on the val data



1. get more training data.
2. Batch norm
3. Data augmentation (fake data)
   crop
4. dropout

2. In homework 3, assume you are implementing a fully connected neural network using the sigmoid activation function. Is this a good idea to initialize the weights with large positive numbers? Explain why or why not.





$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

$$\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$$

x : large

$\sigma(x) \to 0$ or 1

$$\frac{d\sigma(x)}{dx} \to 0.$$

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

$$z = w^T x + b,$$

$$\frac{\partial f(w)}{\partial w} = \frac{\partial z}{\partial w}\frac{\partial f(z)}{\partial z} = \sigma(z)(\underbrace{1-\sigma(z)}_{0})\, X \to 0$$

vanishing gradient = zero gradients = no learning