Terminology

Regression - modeling a continuous function

Classification - modeling a discrete categorical function

Supervised learning - learning a function from (X_i, Y_i) pairs

Unsupervised learning - learning structure from a data set

Overfitting - model learns peculiarities of training set

Underfitting - model insufficiently complex

Sigmoid function - a function that looks like an "S"

Logistic function $\frac{1}{1+e^{-x}}$

Early stopping - stop training when validation error starts to increase

Momentum $V_{t+1} = \mu V_t - \varepsilon \nabla L(\theta_t)$

Weight decay - L2 reg. on weights

Convolutional neural networks (CNNs)

→ used for image/vision tasks

Image (1000×1000)

1	2	1	0	3
4	5	-2	0	2
		4	5	3

filter (2×2)

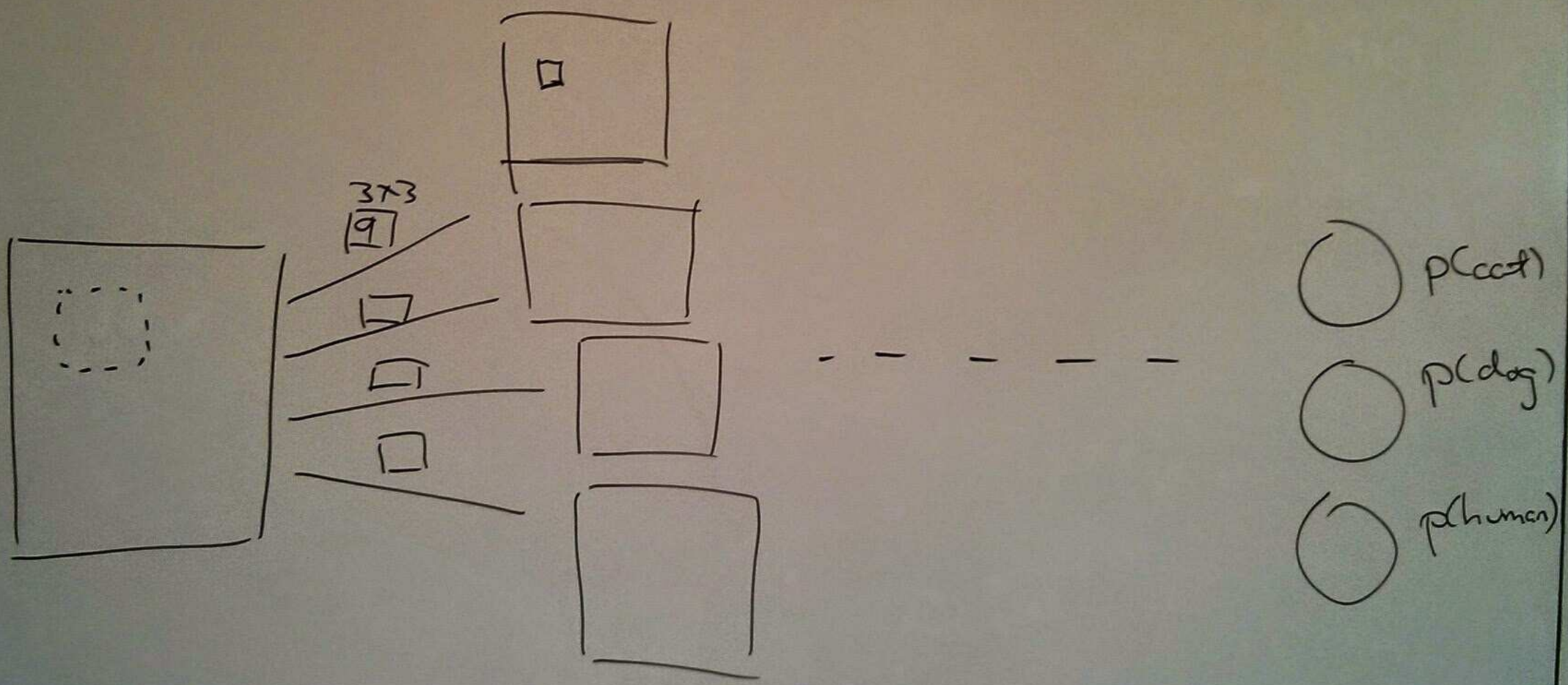
1	-1
1	-1

3	
1	

2	2
-2	-2

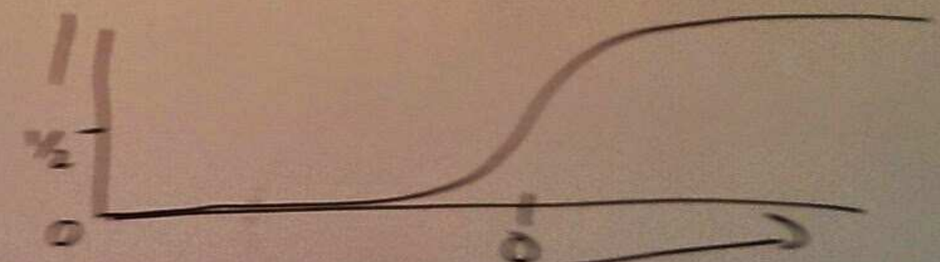
output

5 7 -1

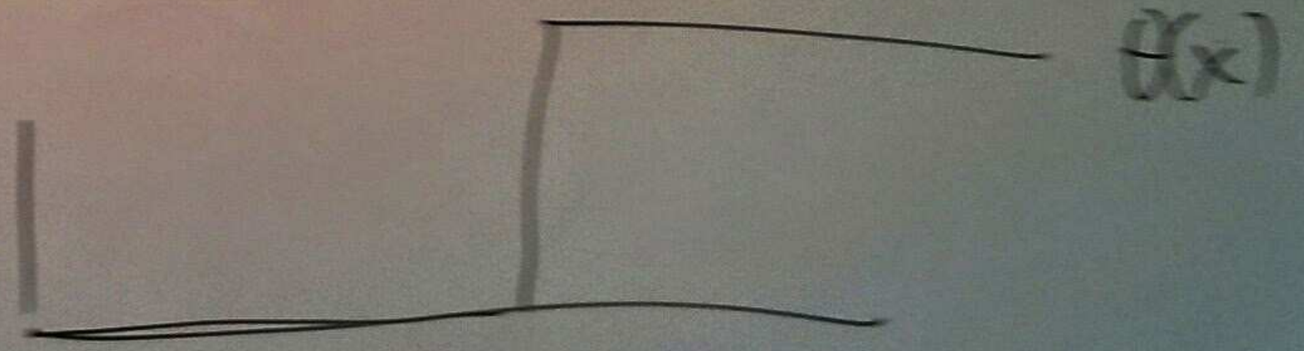

3x3
9
p(cat)
p(dog)
p(human)

Activation functions

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

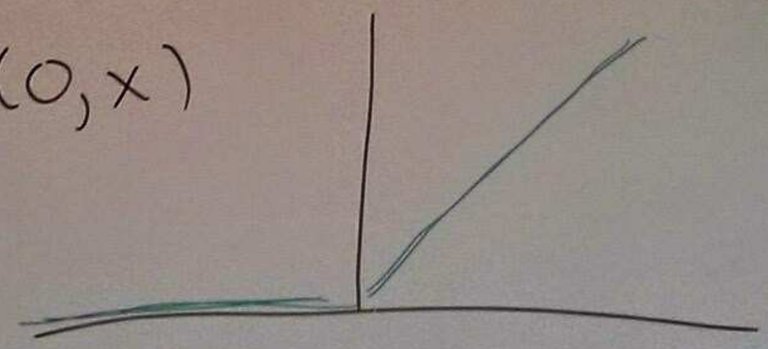


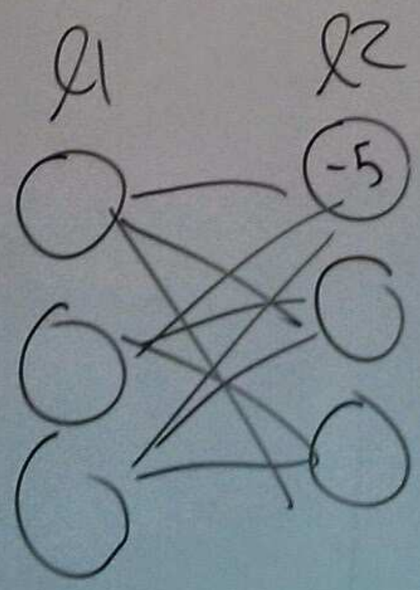
$\sigma(x) = \tanh(x)$



RECtified Linear Units (RELU)

$g_{RELU}(x) = \max(0, x)$

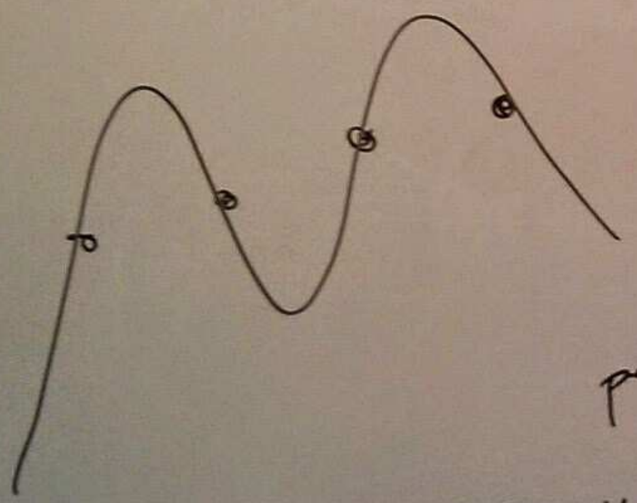


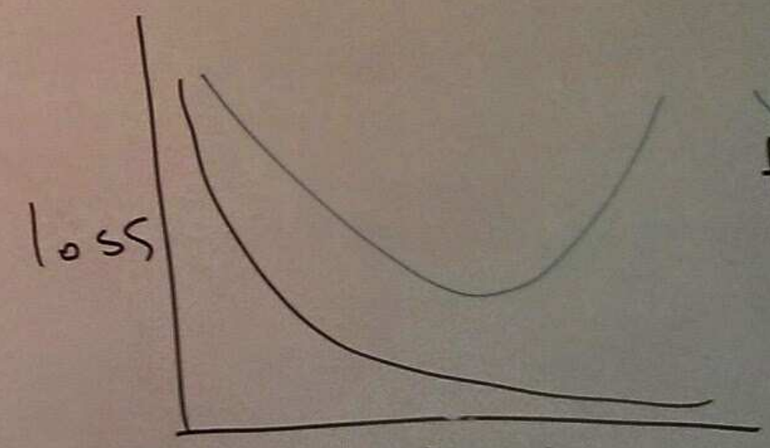


Surveys, Assignments

Overfitting & Regularization

$f(x) = 100 + 259x - 922x^2 - 250x^3$



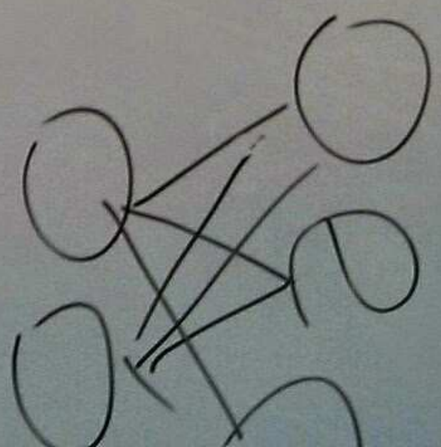


$$L = \sum_j \| \underbrace{y_j}_{\text{predicted output}} - \underbrace{\hat{y}_j}_{\text{true output}} \|_2^2 + \lambda \sum_{\substack{\text{all} \\ \text{weights}}} |w_i|^2$$

Dropout 2012 (a regularization method)

During training, at every iteration, "drop out" (set to 0) each unit with probability p

During prediction, multiply weights by 1-p



Optimization

- Stochastic Gradient Descent

$$W_{t+1} = W_t - \eta \nabla L(W_t)$$

(η: learning rate)

- Momentum

$$W_{t+1} = W_t + V_{t+1}$$

$$V_{t+1} = -\eta \nabla L(W_t) + \varepsilon V_t$$