April 13

Announcements

- Hidden Figures: Tonight 6:30 + through Sunday
- HW3: due date delayed to Monday at noon
  problem 3: shouldn't have more than ~20 clauses

Today Textbook sections 18.4.2, 18.4.3, 18.6, 18.7

Background:

Classification learning problem

Given: $D = (\bar{x}^1, \underset{Y^1}{f(\bar{x}^1)}), (\bar{x}^2, \underset{Y^2}{f(\bar{x}^2)}), \dots, (\bar{x}^m, \underset{Y^m}{f(\bar{x}^m)})$

(each item is indexed by a superscript)

Find: $h(\bar{x}) \approx f(\bar{x})$

$\equiv h(\bar{x})$ with low error

Example of an error function on an example $\bar{x}$

$$E_h(\bar{x}) = (f(\bar{x}) - h(\bar{x}))^2$$ – also called "loss"

$$= \text{Loss}_h(\bar{x})$$

Error of a hypothesis h across all possible $\bar{x}$

$$E_h = \int_x E_h(\bar{x})\, p(\bar{x})\, dx$$ – Error weighted by probability of seeing $\bar{x}$

Empirical error on data

$$\text{Error}_h = \frac{1}{m} \sum_{i=1}^{m} E_h(\bar{x}^i)$$

Common approach to classifier learning:

Find an h with minimal empirical error

$$\text{argmin}_h \text{ Error}_h$$

Returning to perceptrons

Linear Separability

A set of data is linearly separable if
$\exists \bar{w}$ such that for all i $h_{\bar{w}}(\bar{x}^i) = f(\bar{x}^i)$
(or, stated in terms of empirical error
$\exists \bar{w}$ such that $Error_{h_{\bar{w}}} = 0$)

For linear classifiers we'll write $Error_{\bar{w}}$ rather than $Error_{h_{\bar{w}}}$.

Surprising feature of perceptrons
1. Representation analogous to a neuron
2. Learns similar to a neuron
3. Can prove it learns (sometimes)!

Perceptron Convergence Theorem

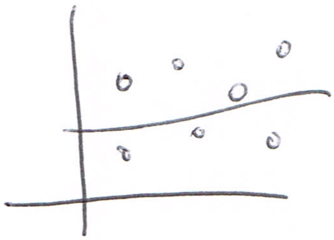
If there is a $\bar{w}$ such that $Error_{\bar{w}} = 0$
then there is an α such that the perceptron learning rule will find a $\bar{w}'$ for which $Error_{\bar{w}'} = 0$

[For <u>any</u> α, if α decays as $O(\frac{1}{t})$ where t is how many updates have taken place - for example, $\frac{1000}{1000+t}$ - then the perceptron will converge]

Recall that the update rule is

$$w_i \leftarrow w_i + \alpha x (f(x) - h(x))$$

Momentary digression to linear regression...



Formula for a line:

$$f(x) = w_1 x + w_0$$

Regression:

Given $(x^1, y^1), \ldots, (x^m, y^m)$

Find w_0, w_1 that minimizes

$$\frac{1}{m} \sum_{i=1}^{m} \left(y^i - (w_1 x^i + w_0)\right)^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left(y^i - h_w(x)\right)^2$$

Solution:

$$w_1 = \frac{m\left(\sum x^i y^i\right) - \left(\sum x^i\right)\left(\sum y^i\right)}{m\left(\sum x^{i^2}\right) - \left(\sum x^i\right)^2}$$

$$w_0 = \frac{\sum y^i - w_1 \sum x^i}{m}$$

Can find the w_0 + w_1 without search

Generalizes to $\bar{x} = (x_1, \ldots, x_n)$ $\bar{w} = (w_0, w_1, \ldots, w_n)$

Can also learn them from the data using <u>gradient descent</u> on the error function

Repeat

For $j = 0$ to n <update each weight by a small amount>

$$w_j \leftarrow w_j - \alpha \frac{\partial \, Loss_w}{\partial w_j}$$

(Recall that Loss is another term for Error)

Until <stopping criterion>

For linear regression

$$Loss_{\bar{w}}(\bar{x}) = (f(\bar{x}) - h_{\bar{w}}(x))^2$$

and the update rule simplifies to

$$w_j \leftarrow w_j - \alpha \sum_{i=1}^{m} \bar{x}^i (f(\bar{x}^i) - h_{\bar{w}}(\bar{x}^i))$$

This is a "batch" update rule, in that it computes the error on <u>all</u> data for each update.

Can instead do this incrementally, updating weights on a per example basis

Known as <u>Stochastic Gradient Descent</u>

Repeat
  For $i=1$ to m <iterate over the data>
    For $j=0$ to n <iterate over the features>

$$w_j \leftarrow w_j - \alpha \frac{\partial Loss_{\bar{w}}(\bar{x}^i)}{\partial w_j}$$

Until <stopping criterion>

[Typically reorder the data each time]

If $Loss_{\bar{w}}(\bar{x}) = (f(\bar{x}) - h_{\bar{w}}(\bar{x}))^2$

the update simplifies to

$$w_j \leftarrow w_j - \alpha x_j^i (f(\bar{x}^i) - h_{\bar{w}}(\bar{x}^i))$$

Same as the Perceptron Learning Rule

Stochastic Gradient Descent is one of the key methods in many forms of deep learning today.

Perceptron problems

1. Problem: Can't represent XOR and other functions

   Solution: Multilayer networks

2. Problem: Unclear how to train lower layers

   Solution: Approximate threshold units

$$h_{\bar{w}}(\bar{x}) = \frac{1}{1 + e^{-\bar{w} \cdot \bar{x}}}$$