Linear Regression
Reminders

• Homework 2: Decision Trees
  – Out: Wed, Jan. 22
  – Due: Wed, Feb. 05 at 11:59pm

• Homework 3: KNN, Perceptron, Lin.Reg.
  – Out: Wed, Feb. 05 (+ 1 day)
  – Due: Wed, Feb. 12 at 11:59pm

• Today’s In-Class Poll
  – http://p7.mlcourse.org
THE PERCEPTRON ALGORITHM
Q: Why do we need an intercept term?

A: It shifts the decision boundary off the origin

Q: Why do we add / subtract 1.0 to the intercept term during Perceptron training?

A: Two cases
1. Increasing b shifts the decision boundary towards the negative side
2. Decreasing b shifts the decision boundary towards the positive side
Perceptron Inductive Bias

1. Decision boundary should be linear
2. Most recent mistakes are most important (and should be corrected)
Background: Hyperplanes

**Notation Trick:** fold the bias $b$ and the weights $w$ into a single vector $\theta$ by prepending a constant to $x$ and increasing dimensionality by one to get $x'$!

Hyperplane (Definition 1):
$$\mathcal{H} = \{x : w^T x = b\}$$

Hyperplane (Definition 2):
$$\mathcal{H} = \{x' : \theta^T x' = 0 \text{ and } x'_1 = 1\}$$
$$\theta = [b, w_1, \ldots, w_M]^T$$

Half-spaces:
$$\mathcal{H}^+ = \{x : \theta^T x > 0 \text{ and } x_0 = 1\}$$
$$\mathcal{H}^- = \{x : \theta^T x < 0 \text{ and } x_0 = 1\}$$
**Data:** Inputs are continuous vectors of length $M$. Outputs are discrete. 

$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots$

where $x \in \mathbb{R}^M$ and $y \in \{+1, -1\}$

**Prediction:** Output determined by hyperplane.

$$\hat{y} = h_\theta(x) = \text{sign}(\theta^T x)$$

Assume $\theta = [b, w_1, \ldots, w_M]^T$ and $x_0 = 1$

**Learning:** Iterative procedure:

- **initialize** parameters to vector of all zeroes
- **while** not converged
  - **receive** next example $(x^{(i)}, y^{(i)})$
  - **predict** $y' = h(x^{(i)})$
  - **if** positive mistake: **add** $x^{(i)}$ to parameters
  - **if** negative mistake: **subtract** $x^{(i)}$ from parameters
(Online) Perceptron Algorithm

**Data:** Inputs are continuous vectors of length $M$. Outputs are discrete.

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots$$

where $x \in \mathbb{R}^M$ and $y \in \{+1, -1\}$

**Prediction:** Output determined by hyperplane.

$\hat{y} = h_\theta(x) = \text{sign}(\theta^T x)$

Assume $\theta = [b, w_1, \ldots, w_M]$

**Implementation Trick:** same behavior as our “add on positive mistake and subtract on negative mistake” version, because $y^{(i)}$ takes care of the sign

**Learning:**

**Algorithm 1** Perceptron Learning Algorithm

1: procedure PERCEPTRON($D = \{(x^{(i)}, y^{(i)})\}$)
2: $\theta \leftarrow 0$
3: for $i \in \{1, 2, \ldots\}$ do
4: $\hat{y} \leftarrow \text{sign}(\theta^T x^{(i)})$
5: if $\hat{y} \neq y^{(i)}$ then
6: $\theta \leftarrow \theta + y^{(i)} x^{(i)}$
7: return $\theta$
Learning for Perceptron also works if we have a fixed training dataset, D. We call this the “batch” setting in contrast to the “online” setting that we’ve discussed so far.

**Algorithm 1** Perceptron Learning Algorithm (Batch)

1: procedure PERCEPTRON($\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$)
2: $\theta \leftarrow 0$ ▷ Initialize parameters
3: while not converged do
4: for $i \in \{1, 2, \ldots, N\}$ do ▷ For each example
5: $\hat{y} \leftarrow \text{sign}(\theta^T x^{(i)})$ ▷ Predict
6: if $\hat{y} \neq y^{(i)}$ then ▷ If mistake
7: $\theta \leftarrow \theta + y^{(i)} x^{(i)}$ ▷ Update parameters
8: return $\theta$
Learning for Perceptron also works if we have a fixed training dataset, D. We call this the “batch” setting in contrast to the “online” setting that we’ve discussed so far.

**Discussion:**
The Batch Perceptron Algorithm can be derived in two ways.

1. By extending the online Perceptron algorithm to the batch setting (as mentioned above)
2. By applying **Stochastic Gradient Descent (SGD)** to minimize a so-called **Hinge Loss** on a linear separator
Extensions of Perceptron

- **Voted Perceptron**
  - generalizes better than (standard) perceptron
  - memory intensive (keeps around every weight vector seen during training, so each one can vote)

- **Averaged Perceptron**
  - empirically similar performance to voted perceptron
  - can be implemented in a memory efficient way (running averages are efficient)

- **Kernel Perceptron**
  - Choose a kernel $K(x', x)$
  - Apply the kernel trick to Perceptron
  - Resulting algorithm is still very simple

- **Structured Perceptron**
  - Basic idea can also be applied when $y$ ranges over an exponentially large set
  - Mistake bound does not depend on the size of that set
The parameter vector \( w \) learned by the Perceptron algorithm can be written as a linear combination of the feature vectors \( x^{(1)}, x^{(2)}, \ldots, x^{(N)} \).

A. True, if you replace “linear” with “polynomial” above

B. True, for all datasets

C. False, for all datasets

D. True, but only for certain datasets

E. False, but only for certain datasets
ANALYSIS OF PERCEPTRON
**Geometric Margin**

**Definition:** The *margin* of example \( x \) w.r.t. a linear sep. \( w \) is the distance from \( x \) to the plane \( w \cdot x = 0 \) (or the negative if on wrong side)
Geometric Margin

Definition: The margin of example $x$ w.r.t. a linear separator $w$ is the distance from $x$ to the plane $w \cdot x = 0$ (or the negative if on wrong side).

Definition: The margin $\gamma_w$ of a set of examples $S$ w.r.t. a linear separator $w$ is the smallest margin over points $x \in S$. 
**Geometric Margin**

**Definition:** The margin of example $x$ w.r.t. a linear sep. $w$ is the distance from $x$ to the plane $w \cdot x = 0$ (or the negative if on wrong side).

**Definition:** The margin $\gamma_w$ of a set of examples $S$ wrt a linear separator $w$ is the smallest margin over points $x \in S$.

**Definition:** The margin $\gamma$ of a set of examples $S$ is the maximum $\gamma_w$ over all linear separators $w$. 

Slide from Nina Balcan
**Def:** For a *binary classification* problem, a set of examples $S$ is *linearly separable* if there exists a linear decision boundary that can separate the points.

<table>
<thead>
<tr>
<th>Case 1:</th>
<th>Case 2:</th>
<th>Case 3:</th>
<th>Case 4:</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Case 1" /></td>
<td><img src="image2.png" alt="Case 2" /></td>
<td><img src="image3.png" alt="Case 3" /></td>
<td><img src="image4.png" alt="Case 4" /></td>
</tr>
</tbody>
</table>

where '+' and '-' represent positive and negative classes, respectively.
Analysis: Perceptron

Perceptron Mistake Bound

**Guarantee:** If data has margin $\gamma$ and all points inside a ball of radius $R$, then Perceptron makes $\leq \left(\frac{R}{\gamma}\right)^2$ mistakes.

(Normalized margin: multiplying all points by 100, or dividing all points by 100, doesn’t change the number of mistakes; algo is invariant to scaling.)
Analysis: Perceptron

**Perceptron Mistake Bound**

**Guarantee:** If data has margin $\gamma$ and all points inside a ball of radius $R$, then Perceptron makes $\leq \left(\frac{R}{\gamma}\right)^2$ mistakes.

(Normalized margin: multiplying all points by 100, or dividing all points by 100, doesn’t change the number of mistakes; algo is invariant to scaling.)

**Def:** We say that the (batch) perceptron algorithm has **converged** if it stops making mistakes on the training data (perfectly classifies the training data).

**Main Takeaway:** For **linearly separable** data, if the perceptron algorithm cycles repeatedly through the data, it will **converge** in a finite # of steps.
Analysis: Perceptron

Perceptron Mistake Bound

**Theorem 0.1** (Block (1962), Novikoff (1962)).

*Given dataset: \( \mathcal{D} = \{ (x^{(i)}, y^{(i)}) \}_{i=1}^{N} \).*

*Suppose:*

1. **Finite size inputs:** \( ||x^{(i)}|| \leq R \)
2. **Linearly separable data:** \( \exists \theta^* \text{ s.t. } ||\theta^*|| = 1 \text{ and } y^{(i)}(\theta^* \cdot x^{(i)}) \geq \gamma, \forall i \)

*Then: The number of mistakes made by the Perceptron algorithm on this dataset is*

\[
k \leq \left( \frac{R}{\gamma} \right)^2
\]
Analysis: Perceptron

Theorem 0.1 (Block (1962), Novikoff (1962)). Given dataset: \( D = \{(x^{(i)}, y^{(i)})\}_{i=1}^N \). Suppose:

1. Finite size inputs: \( \|x^{(i)}\| \leq R \)
2. Linearly separable data: \( \exists \theta^* \text{ s.t. } \|\theta^*\| = 1 \text{ and } y^{(i)}(\theta^* \cdot x^{(i)}) \geq \gamma, \forall i \)

Then: The number of mistakes made by the Perceptron algorithm on this dataset is

\[
k \leq \left(\frac{R}{\gamma}\right)^2
\]

Common Misunderstanding:
The radius is centered at the origin, not at the center of the points.

Figure from Nina Balcan
Analysis: Perceptron

Proof of Perceptron Mistake Bound:

We will show that there exist constants $A$ and $B$ s.t.

$$Ak \leq \|\theta^{(k+1)}\| \leq B\sqrt{k}$$
Theorem 0.1 (Block (1962), Novikoff (1962)).

Given dataset: \( D = \{ (x^{(i)}, y^{(i)}) \}_{i=1}^{N} \).

Suppose:

1. Finite size inputs: \( \| x^{(i)} \| \leq R \)
2. Linearly separable data: \( \exists \theta^* \) s.t. \( \| \theta^* \| = 1 \) and \( y^{(i)} (\theta^* \cdot x^{(i)}) \geq \gamma, \forall i \)

Then: The number of mistakes made by the Perceptron algorithm on this dataset is

\[ k \leq \left( \frac{R}{\gamma} \right)^2 \]

Algorithm 1 Perceptron Learning Algorithm (Online)

```plaintext
1: procedure PERCEPTRON(\( D = \{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots \} \))
2: \( \theta \leftarrow 0, k = 1 \) ▷ Initialize parameters
3: for \( i \in \{1, 2, \ldots \} \) do ▷ For each example
4: if \( y^{(i)} (\theta^{(k)} \cdot x^{(i)}) \leq 0 \) then ▷ If mistake
5: \( \theta^{(k+1)} \leftarrow \theta^{(k)} + y^{(i)} x^{(i)} \) ▷ Update parameters
6: \( k \leftarrow k + 1 \)
7: return \( \theta \)
```
Proof of Perceptron Mistake Bound:
Part 1: for some $A$, $A k \leq \| \theta^{(k+1)} \|

\theta^{(k+1)} \cdot \theta^* = (\theta^{(k)} + y^{(i)} x^{(i)}) \theta^*

by Perceptron algorithm update

$= \theta^{(k)} \cdot \theta^* + y^{(i)} (\theta^* \cdot x^{(i)})$

$\geq \theta^{(k)} \cdot \theta^* + \gamma$

by assumption

$\Rightarrow \theta^{(k+1)} \cdot \theta^* \geq k \gamma$

by induction on $k$ since $\theta^{(1)} = 0$

$\Rightarrow \| \theta^{(k+1)} \| \geq k \gamma$

since $\|w\| \times \|u\| \geq w \cdot u$ and $\|\theta^*\| = 1$

Cauchy-Schwartz inequality
Proof of Perceptron Mistake Bound:
Part 2: for some $B$, $||\theta^{(k+1)}|| \leq B\sqrt{k}$

\[
||\theta^{(k+1)}||^2 = ||\theta^{(k)} + y(i)x(i)||^2
\]

by Perceptron algorithm update

\[
= ||\theta^{(k)}||^2 + (y(i))^2||x(i)||^2 + 2y(i)(\theta^{(k)} \cdot x(i))
\]

\[
\leq ||\theta^{(k)}||^2 + (y(i))^2||x(i)||^2
\]

since $k$th mistake $\Rightarrow y(i)(\theta^{(k)} \cdot x(i)) \leq 0$

\[
= ||\theta^{(k)}||^2 + R^2
\]

since $(y(i))^2||x(i)||^2 = ||x(i)||^2 = R^2$ by assumption and $(y(i))^2 = 1$

\[
\Rightarrow ||\theta^{(k+1)}||^2 \leq kR^2
\]

by induction on $k$ since $(\theta^{(1)})^2 = 0$

\[
\Rightarrow ||\theta^{(k+1)}|| \leq \sqrt{kR}
\]
Analysis: Perceptron

Proof of Perceptron Mistake Bound:
Part 3: Combining the bounds finishes the proof.

\[ k \gamma \leq \| \theta^{(k+1)} \| \leq \sqrt{kR} \]

\[ \Rightarrow k \leq \left( \frac{R}{\gamma} \right)^2 \]

The total number of mistakes must be less than this
Analysis: Perceptron

What if the data is not linearly separable?

1. Perceptron will not converge in this case (it can’t!)
2. However, Freund & Schapire (1999) show that by projecting the points (hypothetically) into a higher dimensional space, we can achieve a similar bound on the number of mistakes made on one pass through the sequence of examples.

**Theorem 2.** Let $\langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ be a sequence of labeled examples with $\|x_i\| \leq R$. Let $u$ be any vector with $\|u\| = 1$ and let $\gamma > 0$. Define the deviation of each example as

$$d_i = \max\{0, \gamma - y_i(u \cdot x_i)\},$$

and define $D = \sqrt{\sum_{i=1}^{m} d_i^2}$. Then the number of mistakes of the online perceptron algorithm on this sequence is bounded by

$$\left( \frac{R + D}{\gamma} \right)^2.$$
Question:
Unlike Decision Trees and K-Nearest Neighbors, the Perceptron algorithm does not suffer from overfitting because it does not have any hyperparameters that could be over-tuned on the training data.

A. True
B. False
C. True and False
Summary: Perceptron

• Perceptron is a **linear classifier**

• **Simple learning algorithm**: when a mistake is made, add / subtract the features

• Perceptron will converge if the data are **linearly separable**, it will **not** converge if the data are **linearly inseparable**

• For linearly separable and inseparable data, we can **bound the number of mistakes** (geometric argument)

• **Extensions** support nonlinear separators and structured prediction
Perceptron Learning Objectives

You should be able to...

• Explain the difference between online learning and batch learning
• Implement the perceptron algorithm for binary classification [CIML]
• Determine whether the perceptron algorithm will converge based on properties of the dataset, and the limitations of the convergence guarantees
• Describe the inductive bias of perceptron and the limitations of linear models
• Draw the decision boundary of a linear model
• Identify whether a dataset is linearly separable or not
• Defend the use of a bias term in perceptron
REGRESSION
Regression

Goal:
- Given a training dataset of pairs \((x,y)\) where
  - \(x\) is a vector
  - \(y\) is a scalar
- Learn a function (aka. curve or line) \(y' = h(x)\) that best fits the training data

Example Applications:
- Stock price prediction
- Forecasting epidemics
- Speech synthesis
- Generation of images (e.g. Deep Dream)
- Predicting the number of tourists on Machu Picchu on a given day
Regression

Example Application: Forecasting Epidemics

• Input features, \( x \): attributes of the epidemic

• Output, \( y \): Weighted %ILI, prevalence of the disease

• Setting: observe past prevalence to predict future prevalence

Figure from Brooks et al. (2015)

Fig 2. 2013–2014 national forecast, retrospectively, using the final revisions of wILI values, using revised wILI data through epidemiological weeks (A) 47, (B) 51, (C) 1, and (D) 7.
Regression

Example: Dataset with only one feature $x$ and one scalar output $y$

Q: What is the function that best fits these points?
**k-NN Regression**

**Example:** Dataset with only one feature $x$ and one scalar output $y$

---

**k=1 Nearest Neighbor Regression**
- **Train:** store all $(x, y)$ pairs
- **Predict:** pick the nearest $x$ in training data and return its $y$

---

**k=2 Nearest Neighbor Distance Weighted Regression**
- **Train:** store all $(x, y)$ pairs
- **Predict:** pick the nearest two instances $x^{(n1)}$ and $x^{(n2)}$ in training data and return the weighted average of their $y$ values
LINEAR REGRESSION
Regression Problems

Chalkboard

– Definition of Regression
– Linear functions
– Residuals
– Notation trick: fold in the intercept