# Assignment8 – Learning

**Given: June 26 Due: July 1**

**Problem 8.1 (Decision List)**

Construct a decision list to classify the data below. The tests should be as small as possible (in terms of attributes), breaking ties among tests with the same number of attributes by selecting the one that classifies the greatest number of examples correctly. If multiple tests have the same number of attributes and classify the same number of examples, then break the tie using attributes with lower index numbers (e.g., select $A_1$ over $A_2$).

| Example | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $y$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 0 | 0 | 0 | 1 |
| $x_2$ | 1 | 0 | 1 | 1 | 1 |
| $x_3$ | 0 | 1 | 0 | 0 | 1 |
| $x_4$ | 0 | 1 | 1 | 0 | 0 |

---

*Solution:* if $A1 = 1$ then 1
else if $A3 = 1$ then 0
else 1

---

**Problem 8.2 (General Properties of Linear Regression)**

Consider a list of examples $(\vec{x}_i, y_i) \in \mathbb{R}^n \times \mathbb{R}$ for $i = 1, \ldots, m$. We want to apply linear regression.

1. What is the hypothesis space?

   ---

   *Solution:* The set of linear functions $\vec{x} \mapsto \vec{w} \cdot \vec{x} + w_0$.

   ---

2. What is the point of the trick to set $(x_i)_0 = 1$ for all examples?

   ---

   *Solution:* If we assume all $x_i$ are of the form $\langle 1, (\vec{x}_i)_1, \ldots, (\vec{x}_i)_m \rangle$ we can write the hypotheses as functions $\vec{x} \mapsto \vec{w} \cdot \vec{x}$ for vectors $\vec{w} \in \mathbb{R}^{1+n}$. That simplifies the calculations.

   ---

3. What is the maximum number of examples for which a consistent hypothesis can still exist?

   ---

   *Solution:* Trick question: There can be infinitely many $\vec{x}$. So if we only

pick examples of the form $(\vec{x}, h(\vec{x}))$ for some fixed hypothesis $h$, we can find infinitely many examples and still have $h$ as a consistent hypothesis.
But we do have the following: If the examples $(\vec{x}, y)$ are chosen randomly and IID, then requiring consistency, i.e., $\vec{w} \cdot \vec{x}_i = y_i$ for $i = 1, \ldots, m$, yields a system of $m$ linear equations in $1 + n$ unknowns (the components of $\vec{w}$). So we can realistically expect being able to find a consistent hypothesis only for up to $1 + n$ examples.

---

4. How important is it to find a consistent hypothesis here?

---

*Solution:* With continuous data (real numbers instead of, e.g., Boolean classifications), a very limited hypothesis space (only linear functions instead of all functions), and many examples (typically much more than $n$), it is unlikely that any hypothesis is consistent. Therefore, practical applications aim for error minimization instead of consistency. A consistent hypothesis would have error 0. Error minimization can be seen as finding the least inconsistent hypothesis.

---

5. What kind of loss function should we use here?

---

*Solution:* $L_{0/1}$ loss (error rate) is not a good choice. It just counts how often $h(\vec{x}_i) \neq y_i$, but equality comparisons on real number data are typically not reliable.
Instead, we should a loss function that measures how far off a hypothesis is. Typical choices are $|h(\vec{x}_i) - y_i|$ and $(h(\vec{x}_i) - y_i)^2$.

---

**Problem 8.3 (Support Vectors)**

Consider the following 2-dimensional dataset

| *support vector* | *classification* |
|---|---|
| $\mathbf{x}_1 = \langle 0, 0 \rangle$ | $\mathbf{y}_1 = -1$ |
| $\mathbf{x}_2 = \langle 0, 0.5 \rangle$ | $\mathbf{y}_2 = -1$ |
| $\mathbf{x}_3 = \langle 0.5, 0 \rangle$ | $\mathbf{y}_3 = -1$ |
| $\mathbf{x}_4 = \langle 1, 1 \rangle$ | $\mathbf{y}_4 = 1$ |
| $\mathbf{x}_5 = \langle 2, 2 \rangle$ | $\mathbf{y}_5 = -1$ |

1. Give a *linear separator* in the form $h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ for the dataset containing only the examples for $\mathbf{x}_1$ to $\mathbf{x}_4$.

---

*Solution:* Many solutions, e.g., $\mathbf{w} = \langle 1, 1 \rangle$ and $b = -1$.

---

2. Explain informally why no linear separator exists for the full dataset of all 5 vectors.

---

*Solution:* The points $\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5$ lie on a line and the middle one has a different classification than the others. No line can have $\mathbf{x}_1$ and $\mathbf{x}_5$ on one side and $\mathbf{x}_4$ on the other.

---

3. Transform the dataset into a 3-dimensional dataset by applying the function $F(\langle u, v\rangle) = \langle u^2, v^2, u + v\rangle$.

---

*Solution:*

| *support vector* $\mathbf{x}$ | $F(\mathbf{x})$ | *classification* |
|---|---|---|
| $\mathbf{x}_1$ | $\langle 0, 0, 0\rangle$ | $\mathbf{y}_1 = -1$ |
| $\mathbf{x}_2$ | $\langle 0, 0.25, 0.5\rangle$ | $\mathbf{y}_2 = -1$ |
| $\mathbf{x}_3$ | $\langle 0.25, 0, 0.5\rangle$ | $\mathbf{y}_3 = -1$ |
| $\mathbf{x}_4$ | $\langle 1, 1, 2\rangle$ | $\mathbf{y}_4 = 1$ |
| $\mathbf{x}_5$ | $\langle 4, 4, 4\rangle$ | $\mathbf{y}_5 = -1$ |

---

4. Give a *linear separator* for the transformed full dataset in the form $h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$.

---

*Solution:* Many solutions, e.g., $\mathbf{w} = \langle -1, -1, 2\rangle$ and $b = -1$.

---

**Problem 8.4 (Weight Updates)**

We're trying to find a linear separator. Our examples are the set

| Example number | $\mathbf{x}_1$ | $\mathbf{x}_2$ | y |
|---|---|---|---|
| 1 | 2 | 0 | 2 |
| 2 | 3 | 1 | 2 |

Our hypothesis space contains the functions $h_{\mathbf{w}}(\mathbf{x}) = A(\mathbf{w}\cdot\mathbf{x})$ for 2+1-dimensional vectors $\mathbf{w}, \mathbf{x}$ (using the trick $\mathbf{x}_0 = 1$ to allow for the constant term $\mathbf{w}_0$) and some fixed function $A$.

As the initial weights, we use $\mathbf{w}_0 = \mathbf{w}_1 = \mathbf{w}_2 = 0$.

For each of the following cases, iterate the respective weight update rule once for each example (using the examples in the order listed). Use learning rate $\alpha = 1$.

1. Using the threshold function $A(z) = \mathcal{T}(z)$, i.e., $A(z) = 1$ if $z > 0$ and $A(z) = 0$ otherwise. Here we cannot do gradient descent, so we have to use the perceptron learning rule.

---

*Solution:* The update rule is $\mathbf{w}_i \leftarrow \mathbf{w}_i + \alpha(y - h_{\mathbf{w}}(\mathbf{x}))\mathbf{x}_i$. Using the examples, we obtain:

- Example 1: $y - h_{\mathbf{w}}(\mathbf{x}) = 2 - \mathcal{T}((0,0,0)\cdot(1,2,0)) = 2$, i.e., $\mathbf{w}_i \leftarrow \mathbf{w}_i + 1\mathbf{x}_i$. Thus, $\mathbf{w} \leftarrow (2,4,0)$.
- Example 2: $y - h_{\mathbf{w}}(\mathbf{x}) = 2 - \mathcal{T}((2,4,0)\cdot(1,3,1)) = 1$, i.e., $\mathbf{w}_i \leftarrow \mathbf{w}_i + 1\mathbf{x}_i$. Thus, $\mathbf{w} \leftarrow (3,7,1)$.

---

2. Using the logistic function $A(z) = 1/(1+e^{-x})$. Here we use gradient descent.

---

*Solution:* The update rule is $\mathbf{w}_i \leftarrow \mathbf{w}_i + \alpha(y - h_{\mathbf{w}}(\mathbf{x}))h_{\mathbf{w}}(\mathbf{x})(1 - h_{\mathbf{w}}(\mathbf{x}))\mathbf{x}_i$. Using the examples, we obtain:

- Example 1: $h_{\mathbf{w}}(\mathbf{x}) = 1/(1+e^0) = 1/2$, i.e., $\mathbf{w}_i \leftarrow \mathbf{w}_i + 1/4(2 - 1/2)\mathbf{x}_i$. Thus, $\mathbf{w} \leftarrow (3/8, 3/4, 0)$.
- Example 2: $h_{\mathbf{w}}(\mathbf{x}) = 1/(1+e^{-((3/8,3/4,0)\cdot(1,3,1))}) = 1/(1+e^{-21/8})$, i.e., (after rounding) $\mathbf{w}_i \leftarrow \mathbf{w}_i + 0.07\mathbf{x}_i$. Thus, $\mathbf{w} \leftarrow (0.44, 0.95, 0.07)$.

---