Learning From Data

Dhruv Rajan

Chapter 1. The Learning Problem

January 2, 2018

Exercise 1.1. Express each task in the framework of *learning from data*. Specify input space $\mathcal{X}$, output space $\mathcal{Y}$, target function $f : \mathcal{X} \to \mathcal{Y}$, and dataset.

1. *Medical diagnosis*
   $\mathcal{X}$: Medical history of past patients. This might include history of specific illnesses, levels of certain chemical, hereditary status for various genes/diseases, etc.
   $\mathcal{Y}$: Diagnoses of past patients

2. *Handwritten Digit Recognition*
   $\mathcal{X}$: Pictures of labeled digits. These may be represented as a vector of pixels, or some condensed feature representation.
   $\mathcal{Y}$: Human generated labels for these digits.

3. *Spam Email Classification*
   $\mathcal{X}$: User emails, spread between spam/ham. Could be represented as hot-vectors of keywords, word counts, etc.
   $\mathcal{Y}$: Classifications of user emails

4. *Predicting how an electric load varies with price, temperature, day of the week.*
   $\mathcal{X}$: Settings for the system, as a 3-vector: $\langle$ price, tempereature, day of week $\rangle$. Should have wide spread of variation between each of these 3 features.
   $\mathcal{Y}$: Measured loads for each setting

Exercise 1.2. Use perceptron to detect spam. Features include frequency of keywords; output $+1$ for spam.

1. Positive Weight
   promotions, medical words, save money, politics.

2. Negative Weight
   Regular words common in non-spam

3. The bias term directly affects how much border-line email gets classified as spam. It serves as a threshold, allowing the separating plane to shift towards conservative of liberal thresholds.

Exercise 1.3. Perceptron Learning Algorithm (PLA) update rule

a) Show that $g = y(t)\mathbf{w}^T\mathbf{x}(t) < 0$

   The hypothesis $h(t)$ is given by $h(t) = \text{sign}(\mathbf{w}^T\mathbf{x}(t))$. For a misclassified point $\mathbf{x}(t)$, we know that $\text{sign}(y(t)) \neq \text{sign}(h(t))$. Thus, the product g mentioned above must be negative, since exactly one of $y(t)$ and $h(t)$ must be negative for some misclassified $\mathbf{x}(t)$.

b) Show that $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^T(t)\mathbf{x}(t)$

$$
\begin{gathered}
y(t)\big[\mathbf{w}(t)+y(t)\mathbf{x}(t)\big]\mathbf{x}(t)\\
y(t)\big[\mathbf{w}(t)\mathbf{x}(t)+y(t)\mathbf{x}^2(t)\big]\\
y(t)\mathbf{w}(t)\mathbf{x}(t)+y^2(t)\mathbf{x}^2(t)
\end{gathered}
$$

Since the factor $y^2(t)\mathbf{x}^2(t)$ is positive, when it is added, it can only increase the initial product $h(t)y(t)$.

c) If the $h(t)\cdot y(t)$ becomes greater than 0, the point $\mathbf{x}(t)$ has become properly classified. Since this product is strictly increased, by the result in (b), it is a step in the right direction.

We can see this geometrically as well.

Exercise 1.4. Classify these situations as either learning or design.

1. Learning
2. Design
3. Learning
4. Design
5. Learning. Though one can optimize analytically for various heuristics, the heuristic has to be picked, and this is a learning problem.

Exercise 1.6. Classify these situations according to their respective learning patterns.

1. Supervised
2. Reinforcement Learning
3. Unsupervised
4. Reinforcement Learning
5. Supervised Learning

Problem 1.1. There are two opaque bags, A, B. A has two black balls, B has 1 black ball, 1 white ball. You pick a bag at random and select a ball from that bag (it is black). What is the probability that the second ball in the bag is also black?

We want P(other ball from same bag is black|first ball is black). Bayes rule gives us that

$$P[A\cap B]=P[A|B]\cdot P[B]=P[B|A]\cdot P[A] \tag{1}$$

$$
\begin{aligned}
P[\text{1st black}|\text{2nd black}] &= P[\text{2nd black}|\text{1st black}]\cdot P[\text{1st black}]\\
P[\text{2nd black}|\text{1st black}] &= \frac{P[\text{1st black}\cap\text{2nd black}]}{P[\text{1st black}]}\\
&= \frac{0.5}{0.75}\\
&= \boxed{\tfrac{2}{3}}
\end{aligned}
$$

Problem 1.2. Consider the two-dimensional perceptron $h(x)=\text{sign}(\mathbf{w}^T\mathbf{x})$.

a) Show that the regions on the plaine where $h(x) = +1$ and $h(x) = -1$ are separated by a line. Express this in slope-intercept form. The hypothesis is given by

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x})$$

The boundary line is given by:

$$\begin{aligned} 0 &= \mathbf{w}^T\mathbf{x} \\ &= w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 \end{aligned}$$

This is a linear equation in two variables. Thus, the boundary must be linear. To find the equation of this line in slope intercept form, we can solve for x_2 in terms of x_1.

$$x_2 = \frac{-w_1}{w_2} \cdot x_1 - \frac{w_0}{w_2}$$

b) Line for $w = [1, 2, 3]$ is $x_2 = -\frac{2}{3} \cdot x_1 - \frac{1}{3}$. Line for $w = [-1, -2, -3]$ is $x_2 = -\frac{2}{3} - \frac{1}{3}$.

Problem 1.3. Prove that the PLA eventually converges to a linear separator for separable data. Assume $\mathbf{w}(0) = 0$.

a) Let $\rho = \min_{1 \le n \le N} y_n(\mathbf{w}^{*T}\mathbf{x}_n)$. Show that $\rho > 0$. We know that $\mathbf{w}^*$ correctly classifies all points. Thus, the sign of the product $\mathbf{w}^{*T}\mathbf{x}_n$ must match the sign of y_n. When any two quantities of the same sign are multiplied, the resulting value is postive, so $\rho > 0$.

b) Show that $\mathbf{w}(t)\mathbf{w}^* \ge \mathbf{w}^T(t-1)\mathbf{w}^* + \rho$, and conclude that $\mathbf{w}(t)\mathbf{w}^* \ge t\rho$.

The update rule is as follows:

$$\mathbf{w}(t) = \mathbf{w}(t-1) + y(t)\mathbf{x}(t)$$

We can use this to reduce the left hand of the inequality:

$$\begin{aligned} \mathbf{w}(t)\mathbf{w}^* &= \mathbf{w}^* \cdot [\mathbf{w}(t-1) + y(t)\mathbf{x}(t)] \\ &= \mathbf{w}^*\mathbf{w}(t-1) + \underbrace{\mathbf{w}^* y(t)\mathbf{x}(t)}_{\text{call this term } s} \end{aligned}$$

The term s must be $\ge \rho$ since ρ specifies the minimum possible value (over all t) of this quantity. Since $s \ge \rho$, the statement must remain true if we substitute ρ for s, and thus, we have the first inequality.

Next, we want to show that $\mathbf{w}(t)\mathbf{w}^* \ge t\rho$. We show this by induction. At time $t = 0$:

$$\mathbf{w}^T \cdot \mathbf{w}^* = 0 \ge 0 \cdot \rho \quad \checkmark$$

Given that $\mathbf{w}(t)\mathbf{w}^* \ge t\rho$, we need to show this holds for time $t+1$. From the previous result, we have that

$$\begin{aligned} \mathbf{w}(t+1)\mathbf{w}^* &\ge \mathbf{w}(t)\mathbf{w}^* + \rho \\ &\ge t\rho + \rho \\ &\ge (t+1)\rho \end{aligned}$$

c) Show that $\|\mathbf{w}(t)\|^2 \le \|\mathbf{w}(t-1)\|^2 + \|\mathbf{x}(t-1)\|^2$.

This is a restatement of the triangle inequality.

d) Show that $\|\mathbf{w}(t)\|^2 \leq tR^2$ where $R = \max_{1 \leq n \leq N} \|x_n\|$ At time $t = 0$

$$\|\mathbf{w}(t)\|^2 = 0 \leq 0 \cdot R^2 \quad \checkmark$$

Given $\|\mathbf{w}(t)\|^2 \leq t \cdot R^2$, we need this to hold for time $t + 1$.

$$\begin{aligned} \|\mathbf{w}(t+1)\|^2 &\leq \|\mathbf{w}(t)\|^2 + \|\mathbf{x}(t)\| \\ &\leq tR^2 + \|\mathbf{x}(t)\| \end{aligned}$$

It is necessarily the case that $R >= \|\mathbf{x}(t)\|$, since R is the maximum (over all t) of this quantity. Thus, we can substitute R or R^2 in place of this expression.

$$\begin{aligned} \|\mathbf{w}(t+1)\|^2 &\leq tR^2 + R^2 \\ &\leq (t+1)R^2 \end{aligned}$$