Learning From Data - worked examples

Mac Radigan

Problem 1.1

We have 2 opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black and a white ball. You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball it is black. You now pick the second ball from that same bag. What is the probability that this ball is also black? [Hint: Use Bayes' Theorem: $P[AandB] = P[A|B]P[B] = P[B|A]P[A]$.]

$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

- B_A the event that bag A was selected
- B_B the event that bag B was selected
- k_A the event that a black ball from bag A was selected
- k_B the event that a black ball from bag B was selected
- k_1 the event that a black ball was selected on the first selection
- k_2 the event that a black ball was selected on the second selection

$k_1 = B_A k_A + B_B k_B = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{4}$

$P(B_A|k_1) = \frac{P(B_A \cap k_1)}{P(k_1)} = \frac{P(k_1|B_B)P(B_A)}{P(k_1)}$

$P(B_B|k_1) = \frac{P(B_B \cap k_1)}{P(k_1)} = \frac{P(k_1|B_B)P(B_B)}{P(k_1)}$

$P(k_1|B_A) = \frac{1}{2} \cdot 1 = \frac{1}{2}$

$P(k_1|B_B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$P(B_A) = \frac{1}{2}$

$P(B_B) = \frac{1}{2}$

$P(k_1) = \frac{3}{4}$

$P(B_A|k_1) = \frac{P(k_1|B_A)P(B_A)}{P(k_1)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$

$P(B_B|k_1) = \frac{P(k_1|B_B)P(B_B)}{P(k_1)} = \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{\frac{1}{8}}{\frac{3}{4}} = \frac{4}{24} = \frac{1}{6}$

$P(k_2) = P(k_1|B_B)P(k_B) + P(B_B|k_1)P(k_B) = \frac{1}{3} \cdot 1 + \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{3} + \frac{1}{12} + \frac{4}{12} + \frac{1}{12} = \frac{5}{12}$

Problem 1.2

Consider the perceptron in two dimensions: $h(x) = sign(w^{\intercal}x)$ where $w = [w_0, w_1, w_2]$ and $x = [1, x_1, x_2]$. Technically, x has three coordinates, but we call this perceptron two-dimensional because the first coordinate is fixed at 1.

(a) Show that the regions on the plane where $h(x) = +1$ and $h(x) = -1$ are separated by a line. If we express this line by the equation $x_2 = ax_1 + b$, what are the slope a and intercept b in terms of w_0, w_1, w_2?

$w^{\intercal}x = 0$

$\implies w_0x_0 + w_1 + x_1 + w_2x_2 = 0$

$\implies w_2x_2 = -w_1x_1 - w_0x_0 = -w_1x_1 - x_0$

$\implies x_2 = -\frac{w_1}{w_2}x_2 - \frac{w_0}{w_2}$

$a = -\frac{w_1}{w_2}$

$b = -\frac{w_0}{w_2}$

(b) Draw a picture for the cases $w = [1, 2, 3]$ and $w = -[1, 2, 3]$.

In more than two dimensions, the +1 and -1 regions are separated by a hyperplane, the generalization of a line.

Problem 1.3

Prove that the PLA eventually converges to a linear separator for separable data. The following steps will guide you through the proof. Let w^* be an optimal set of weights (one which separates the data). The essential idea in this proof is to show that the PLA weights $w(t)$ get "more aligned" with w^* with every iteration. For simplicity, assume that $w(0) = 0$.

(a) Let $\rho = min_{1\leq n\leq N}(w^{*\intercal}x_n)$. Show that $\rho > 0$.

We know $w^{*\intercal}x_n > 0\forall n$, since x is linearly separable, and w^* separates x.

So by definition,

$w^{*\intercal}x_n > 0 \; \forall n$ s.t. $y_n = +1$

and

$w^{*\intercal}x_n < 0 \; \forall n$ s.t. $y_n = -1$

and thus

$y_n \left(w{*}^{\intercal}x_n\right) \forall n$

that is, ρ is strictly positive.

(b) Show that $w^\intercal(t)w^* \geq w^\intercal(t-1)w^* + \rho$, and conclude that $w^\intercal(t) \geq t\rho$.

Let $x^{'}$ be the set of misclassified points, and let $y^{'}$ be the corresponding truth values

Further, let $\rho_m = \underset{m}{min}\ y_m\left(w^{*\intercal}x_m\right)$

Then $w_t = \Sigma_{m=0}^{t} y_m^{'} x_m^{'}$

But since $\left(y_m^{'}x_m^{'}\right) \leq \left(y_t^{'}x_t^{'}\right)^\intercal w^*\ \forall m, t$

then

$\left(\Sigma_{m=0}^{t}y_m^{'}x_m^{'}\right)w^* \geq \left(\Sigma_{m=0}^{t-1}y_m^{'}x_m^{'}\right)^\intercal w^* + y_m\left(x_m^\intercal w^*\right)$

and thus

$w^\intercal(t) \geq t\rho$

(c) Show that $\|w(t)\|^2 \leq \|w(t-1)\| + \|w(t-1)\|^2$.

[Hint: $y(t-1) \cdot (w^\intercal(t-1)x(t-1)) \leq 0$ because $x(t-1)$ was misclassified by $w(t-1)$.]

Note $w_t = w_{t-1} + y_{t-1}x_{t-1}$

so

$\|w_{t-1} + y_{t-1}x_{t-1}\|^2 \leq \left\|w_{t-1}^2\right\|^2 + \left\|x_{t-1}^2\right\|^2$

then

$\left(w_{t-1} + y_{t-1}x_{t-1}\right)^2 \triangleq w_{t-1}^2 + 2w_{t-1}^\intercal y_{t-1}x_{t-1} + y_{t-1}^2x_{t-1}^2$

but $2w_{t-1}^\intercal y_{t-1}x_{t-1} < 0$

thus

$\|w(t)\|^2 \leq \|w(t-1)\| + \|w(t-1)\|^2$

(d) Show by induction that $\|w(t)\|^2 \leq tR^2$, where $R = max_{1\leq n\leq N}\|x_n\|$.

Base case $t = 1$:

$w_1 = w_0 + x_m^{'}y_m^{'} = 0 + x_m^{'}y_m^{'} = y_m^{'}x_m^{'}$, and $\|x_m\| \leq \|x_n\|$

(e) Using (b) and (d), show that

$\frac{w^\intercal}{\|w(t)\|^2}w^* \geq \sqrt{t} \cdot \frac{\rho}{R}$,

and hence prove that

$t \leq \frac{R^2\|w^*\|^2}{\rho^2}$.

[Hint: $\frac{w^\intercal(t)w^*}{\|w(t)\|\|w^*\|} \leq 1$. Why?]

Note that $\frac{w_t^\intercal w^*}{\|w_t\|\|w_t^*\|} \leq 1$

so $\frac{w_t^\intercal w^*}{\|w_t\|} \leq \|w_t^*\|$

Now $\|w_t\|^2 \leq tR^2 \Rightarrow \|w_t\| \leq \sqrt[2]{t}R$

Then $\frac{1}{\|w_t\|} \geq \frac{1}{\sqrt[2]{t}R}$

Thus $\frac{w^\intercal w^*}{\geq} \frac{\sqrt[2]{t}\rho}{R}$

In practice, PLA converges more quickly than the bound $\frac{R^2\|w^*\|^2}{\rho^2}$ suggests. Nevertheless, because we do not know p in advance, we can't determine the number of iterations to convergence, which does pose a problem if the data is non-separable.

Problem 1.4

In Exercise 1.4, we use an artificial data set to study the perceptron learning algorithm. This problem leads you to explore the algorithm further with data sets of different sizes and dimensions.

(a) Generate a linearly separable data set of size 20 as indicated in Exercise 1.4. Plot the examples (x_n, y_n) as well as the target function f on a plane. Be sure to mark the examples from different classes differently, and add labels to the axes of the plot.

(b) Run the perceptron learning algorithm on the data set above. Report the number of updates that the algorithm takes before converging. Plot the examples (x_n, y_n), the target function f, and the final hypothesis g in the same figure. Comment on whether f is close to g.

(c) Repeat everything in (b) with another randomly generated data set of size 20. Compare your results with (b).

(d) Repeat everything in (b) with another randomly generated data set of size 100. Compare your results with (b).

(e) Repeat everything in (b) with another randomly generated data set of size 1, 000. Compare your results with (b).

(f) Modify the algorithm such that it takes $x_n \in \mathbb{R}^n$ instead of $\mathbb{R}^2$. Randomly generate a linearly separable data set of size 1,000 with $x_n \in \mathbb{R}^10$ and feed the data set to the algorithm. How many updates does the algorithm take to converge?

(g) Repeat the algorithm on the same data set as (f) for 100 experiments. In the iterations of each experiment, pick $x(t)$ randomly instead of determin-

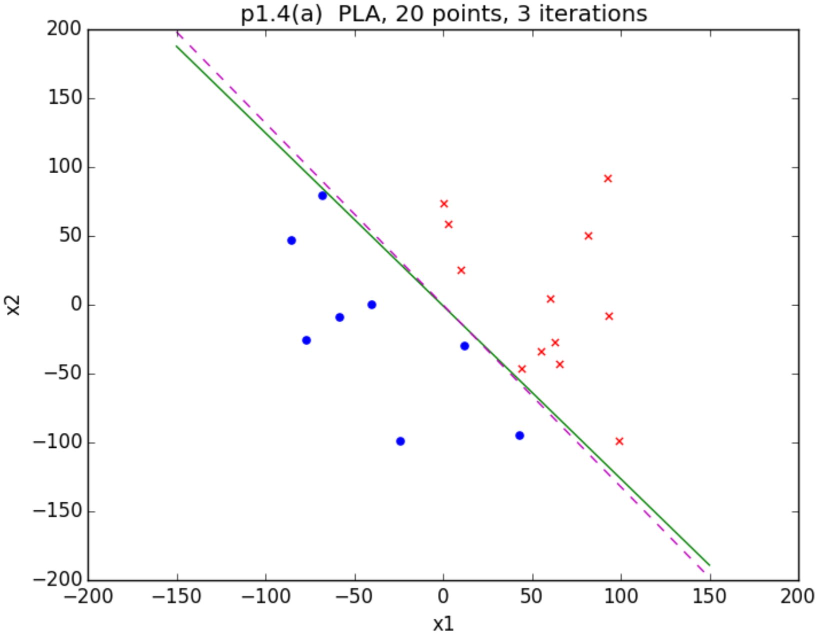

Figure 1: Target

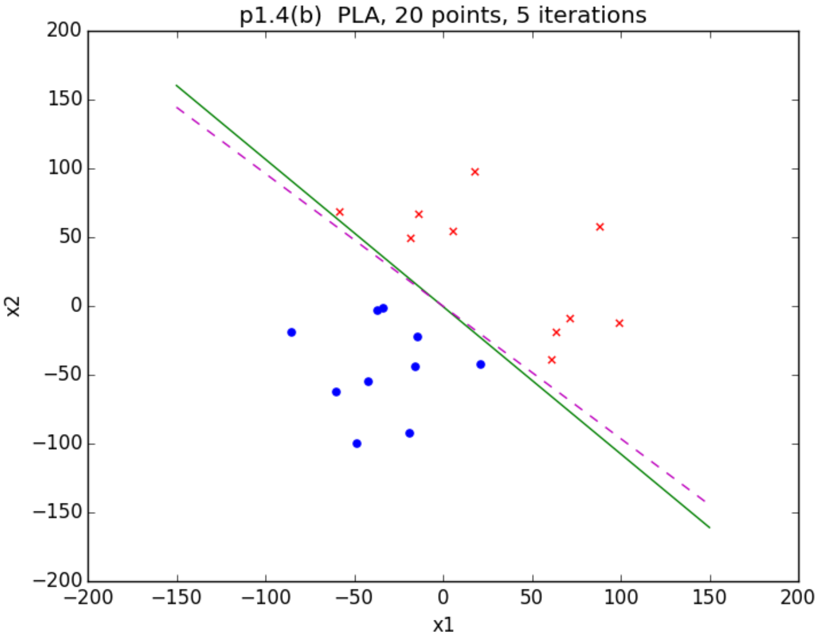

Figure 2: Target

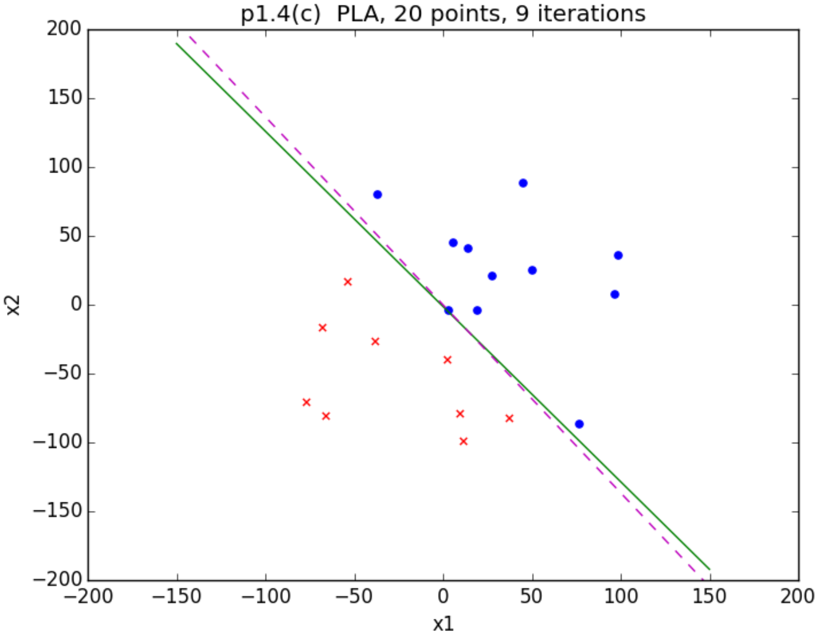

Figure 3: Target

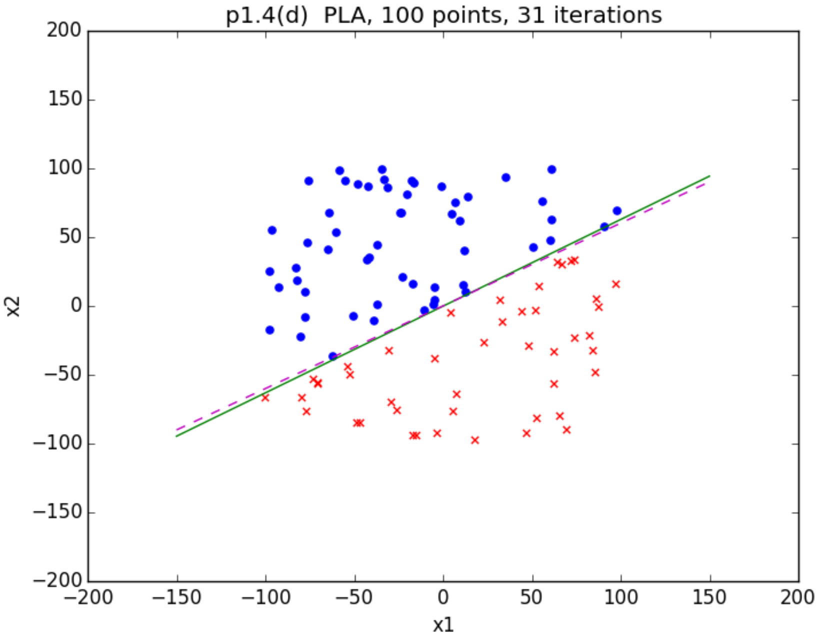

Figure 4: Target

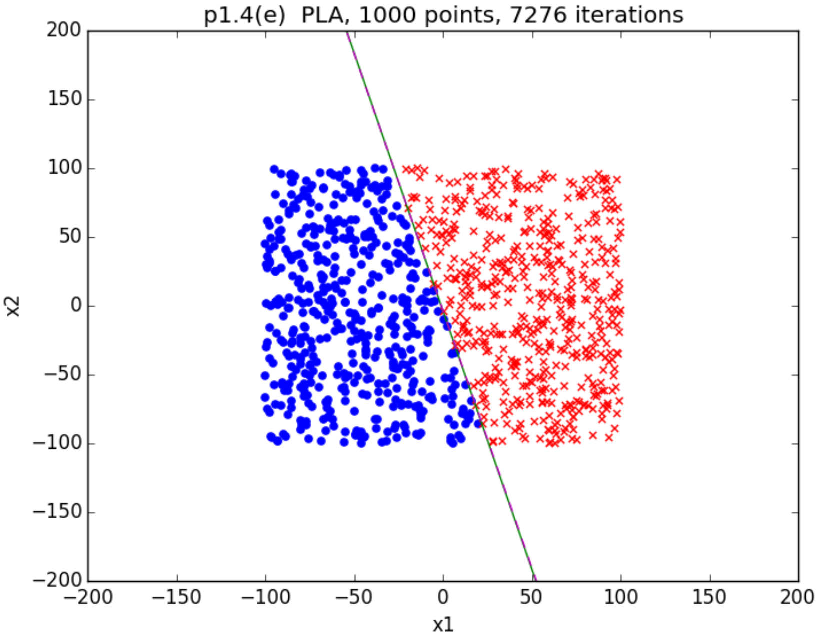

Figure 5: Target

istically. Plot a histogram r the number of updates that the algorithm takes to converge.

(h) Summarize your conclusions with respect to accuracy and running time as a function of N and d.

Problem 1.8

The Hoeffding Inequality is one form of the law of large numbers. One of the simplest forms of that law is the Chebyshev Inequality, which you will prove here.

(a) If t is a non negative random variable, prove that for any $\alpha > 0$, $P[t \geq \alpha] \leq \frac{E(t)}{\alpha}$

By definition,

$P[t \geq \alpha] \triangleq \int_{\alpha}^{\infty} p(x)dx$

and

$E(t) \triangleq \int_{-\infty}^{\infty} xp(x)dx$

so evaluate

$\alpha \int_{\alpha}^{\infty} p(x)dx = \alpha P[t \geq \alpha] \leq E(t) = \int_{-\infty}^{\infty} xp(x)dx$

since t is strictly positive, this can be written as

$\int_{\alpha}^{\infty} \alpha p(x)dx \geq \int_{0}^{\alpha} xp(x)dx + \int_{\alpha}^{\infty} xp(x)dx$

note that $\int_{\alpha}^{\infty} \alpha p(x)dx \geq \int_{\alpha}^{\infty} xp(x)dx$

and because $t > 0$, $\int_{0}^{\alpha} xp(x)dx \geq 0$

so it holds that $\alpha \int_{\alpha}^{\infty} p(x)dx = \int_{-\infty}^{\infty} xp(x)dx$

and thus $P[t \geq \alpha] \leq \frac{E(t)}{\alpha}$

(b) If u is any random variable with mean μ and variance σ^2, prove that for any $\alpha > 0, P[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$. [Hint: Use (a)]

By definition,

$\sigma^2 \triangleq E[(u-\mu)^2]$

and thus with substitution of $t := (u-\mu)^2$ and using (a), we have

$P[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$

(c) If $u_1, \ldots, u_N$ are iid random variables, each with mean μ and variance σ^2, and $u = \frac{1}{N}\Sigma_{n=1}^{N} u_n$, prove that for any $\alpha > 0$, $P[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}$.

Notice that the RHS of this Chebyshev Inequality goes down linearly in N, while the counterpart in Hoeffding's Inequality goes down exponentially. In Problem 1.9, we develop an exponential bound using a similar approach.