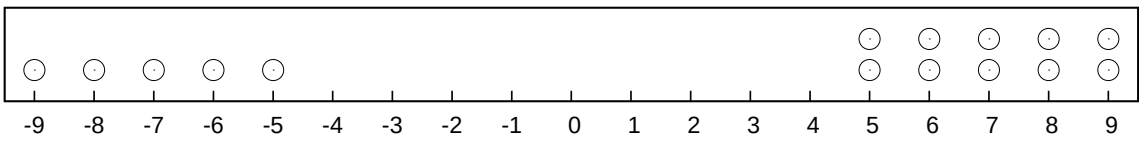
# Introductory Applied Machine Learning, Tutorial 4 Answers

September 2017

1. The maximum margin hyperplane is
$$x_1 + x_2 = 1.$$
This can't be proven with the theoretical tools that we have seen, but intuitively the points $(0, 0.5)$ and $(0.5, 1.0)$ are the closest pair of points from opposite classes, and this plane cuts them halfway and perpendicularly. Any other plane would be closer to one or the other of these points, and therefore have a smaller margin.

   We can write this equivalently as
$$2x_1 + 2x_2 = 2$$
   to satisfy the constraint
$$\min_i |\mathbf{w}^T \mathbf{x}_i + w_0| = 1$$
   that we used to derive the SVM optimization problem. We represent this using the weight vector
$$\mathbf{w} = [2, 2], w_0 = -2$$
   The support vectors are the three points $(0.5, 1), (1, 0.5)$, and $(0, 0.5)$. The margin as we have defined it is the distance from the closest point to the hyperplane, i.e.,
$$\min_i \frac{1}{||\mathbf{w}||} |\mathbf{w}^T \mathbf{x}_i + w_0|.$$
   Using the choice of $\mathbf{w}$ above, the margin is $1/\sqrt{8}$.

2. Here is a diagram of the data




   (a) Intuitively, there are two groups of points, $a = \{-9, \ldots, -5\}$ and $b = \{+5, \ldots, +9\}$.
$$\begin{aligned} \mu_a &= \frac{-9-8-7-6-5}{5} = -7 \\ \mu_b &= \frac{5+5+6+6+7+7+8+8+9+9}{10} = +7 \\ p_a &= \frac{5}{5+10} = \frac{1}{3} \\ p_b &= \frac{2}{3} \end{aligned}$$

(b)

| $x_i$: | -9 | -8 | -7 | -6 | -5 | +5 | +5 | |
|---|---|---|---|---|---|---|---|---|
| $p(x_i \mid a)$: | .24 | .05 | .004 | .0001 | $1.4 \cdot 10^{-6}$ | $10^{-50}$ | $10^{-50}$ | |
| $p(x_i \mid b)$: | $2 \cdot 10^{-27}$ | $2 \cdot 10^{-32}$ | $8 \cdot 10^{-38}$ | $10^{-43}$ | $10^{-50}$ | $10^{-137}$ | $10^{-137}$ | |
| $a_i = p(a \mid x_i)$: | $\approx 1$ | 1 | 1 | 1 | 1 | 1 | 1 | |
| $b_i$: | $8 \cdot 10^{-27}$ | $10^{-31}$ | $10^{-35}$ | $10^{-40}$ | $10^{-44}$ | $10^{-87}$ | $10^{-87}$ | |
| $x_i$: | +6 | +6 | +7 | +7 | +8 | +8 | +9 | +9 |
| $p(x_i \mid a)$: | $10^{-56}$ | $10^{-56}$ | $10^{-64}$ | $10^{-64}$ | $10^{-71}$ | $10^{-71}$ | $10^{-79}$ | $10^{-79}$ |
| $p(x_i \mid b)$: | $10^{-148}$ | $10^{-148}$ | $10^{-159}$ | $10^{-159}$ | $10^{-171}$ | $10^{-171}$ | $10^{-184}$ | $10^{-184}$ |
| $a_i = p(a \mid x_i)$: | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $b_i$: | $10^{-92}$ | $10^{-92}$ | $10^{-96}$ | $10^{-96}$ | $10^{-100}$ | $10^{-100}$ | $10^{-105}$ | $10^{-105}$ |

$p_a \approx 1$ and $p_b \approx 10^{-27}$, so every point is assigned to $a$.

$$
\begin{aligned}
\mu_a &= 2.333 \ldots \text{overall mean (of all points)} \\
\sigma_a &= 6.749 \\
\mu_b &\approx -9 \ldots b_i \text{ are all near zero, but some "zeroes" are much smaller than others} \\
\sigma_b &= 0.006 \ldots \text{all mass of } b \text{ focused on one point}
\end{aligned}
$$

(c) In the next iteration the variance of component $b$ will shrink essentially to zero, which will cause problems. If it is bounded away from zero at some small value then component $b$ will probably "grab" the datapoint at -9, and the others will be covered by component $a$. There is no guarantee that $\mu_a$ and $\mu_b$ will travel to the computed values in part 3.1.