1. (a) Because the points are uniformly distributed, the expected fraction of the points residing on a subinterval is the length of the subinterval compared to the length of the whole interval. If we denote the length of the interval by $\alpha$, the expected number of points falling within the interval, $E(\alpha)$, is given by

$$E(\alpha) = \frac{0.1}{1} = 0.1.$$ 

(b) In two dimensions, we have likewise that the expected number of points is given by the relative area of the smaller square compared to the area of the whole square:

$$E(\alpha) = \frac{0.1^2}{1^2} = 0.01.$$ 

(c) For general $p \geq 1$, the expected fraction is the relative volume of the smaller hypercube compared to the volume of the whole hypercube:

$$E(\alpha) = \frac{0.1^{100}}{1^{100}} = 10^{-100}.$$ 

(d) If we assume the observations are uniformly distributed and use aforementioned definition of "nearness", on average only $10^{-p}$ training observations are near the given test observation when the number of features is $p \geq 1$.

In practice, things may turn out not to be as bad as this: observations are usually not uniformly distributed, and sometimes two observations do not have to be near each other in all $p$ dimensions to be considered similar to each other — for example, some of the features may be irrelevant in view of the task at hand.

(e) We can solve the length, $x$, of the side of the hypercube that contains on the average 10% of the data:

$$\frac{1}{10} = E(\alpha) = \frac{x^p}{1^p}$$

$$x = 10^{-1/p}.$$ 

For $p = 1, 2$ and 100, this is:

$$x = 10^{-1} = 0.1$$
$$x = 10^{-1/2} \approx 0.3162$$
$$x = 10^{-1/100} \approx 0.9772.$$ 

So the hypercube that contains on the average only 10% of the data has sides of length approximately 0.9772, so almost 1! This demonstrates that data is sparse in high-dimensional spaces. In other words, the training observations that are among the 10% of the training data nearest to the test observation may actually be almost maximally different from the test observation.
2. Define a binary random variable $D$ as follows:

$$D = \begin{cases} 
1, & \text{if a company pays a dividend} \\
0, & \text{otherwise.}
\end{cases}$$

The prior probabilities for $D$ are $\Pr(D = 0) = \pi_0 = 0.2$, and $\Pr(D = 1) = \pi_1 = 0.8$.

The conditional distributions of the percentage profit $X$ given values of $D$ are:

$$f_{X \mid D=0}(x) = \mathcal{N}(x; \mu_0, \sigma^2)$$

$$f_{X \mid D=1}(x) = \mathcal{N}(x; \mu_1, \sigma^2)$$

where the expected values are $\mu_0 = 0$, $\mu_1 = 10$, and a common variance is $\sigma^2 = 36$.

We get the probability that a company pays a dividend given that its percentage profit is $X = 4$ using the Bayes’ theorem:

$$P(D = 1 \mid X = 4) = \frac{\pi_1 f_{X \mid D=1}(4)}{\pi_1 f_{X \mid D=1}(4) + \pi_0 f_{X \mid D=0}(4)}$$

$$= \frac{0.8 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(4-\mu_1)^2}{2\sigma^2} \right)}{0.8 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(4-\mu_1)^2}{2\sigma^2} \right) + 0.2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(4-\mu_0)^2}{2\sigma^2} \right)}$$

$$= \frac{0.8e^{-\frac{1}{2}}}{0.8e^{-\frac{1}{2}} + 0.2e^{-\frac{29}{36}}} \approx 0.752$$

Thus, the probability that the company pays a dividend given that its percentage profit is 4.0 is about 75%.

Note that this is just slightly lower than the prior probability, $\Pr(D = 1)$, which was 80%. This is consistent with the reasoning that the average percentage profit of companies that pay dividends, $\mu_1 = 10.0$, is a little further away from the observed profit, $x = 4.0$, than the average profit of companies that don’t pay dividends, $\mu_0 = 0.0$. So while the observed profit is slightly more typical to non-dividend-paying companies, the difference is so small that the effect on the conditional probability is less than 5 percentage points.

We are of course curious to know what the posterior probability would have looked like if the profit percentage, $x$, had been something different. To look into this, let’s write a little R script that computes the posterior $P(D = 1 \mid X = x)$ for different $x$:

```r
po <- function(x) { a=.8*exp(-(x-10)^2/(2*36)); b=.2*exp(-(x-0)^2/(2*36)); a/(a+b) }
x = .5*(-30:50) # a grid of points between -15 and 25 plot(x, po(x)) # basic plot showing the posterior as a function of x plot(x, po(x), t='h', lwd=6, col='orange', lend=2) # fancy plot showing the same points(4, po(4),t='h',lwd=6,col='black',lend=2) # emphasize x=4 in black
```

(plot shown on the next page...)

---

1 Bayes’ theorem applies also for the joint distribution of the discrete and continuous random variable. For a continuous variable probabilities in the formula are replaced by densities, and conditional probabilities by conditional densities.
As you can see, the model predicts that the chances that a company pays dividends decrease as the profit $x$ decreases, and vice versa. What we have managed to implement here is actually one-dimensional **Linear Discriminant Analysis (LDA)**!

In next week’s exercises, we’ll do something similar in two-dimensions.