Week 9-1: Anomaly Detection

1. For which of the following problems would anomaly detection be a suitable algorithm?

   (a) From a large set of primary care patient records, identify individuals who might have unusual health conditions.

   (b) Given a dataset of credit card transactions, identify unusual transactions to flag them as possibly fraudulent.

2. Suppose you have trained an anomaly detection system for fraud detection and your system flags anomalies when $p(x) < \epsilon$, and you find on the cross-validation set that is mis-flagging far too many good transactions as fraudulent. What should you do?

   Decrease ϵ. The threshold is then decreased for fewer flags.

3. Suppose you are developing an anomaly detection system to catch manufacturing defects in airplane engines. Your model uses $p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$. You have two features x_1 = vibration intensity and x_2 = heat generated. Both x_1 and x_2 take on values between 0 and 1 (and are strictly greater than 0), and for most "normal" engines you expect that $x_1 \approx x_2$. One of the suspected anomalies is that a flawed engine may vibrate very intensely even without generating much heat (large x_1, small x_2) even though the particular values of x_1 and x_2 may not fall outside their typical range of values. What additional feature x_3 should you create to capture these types of anomalies?

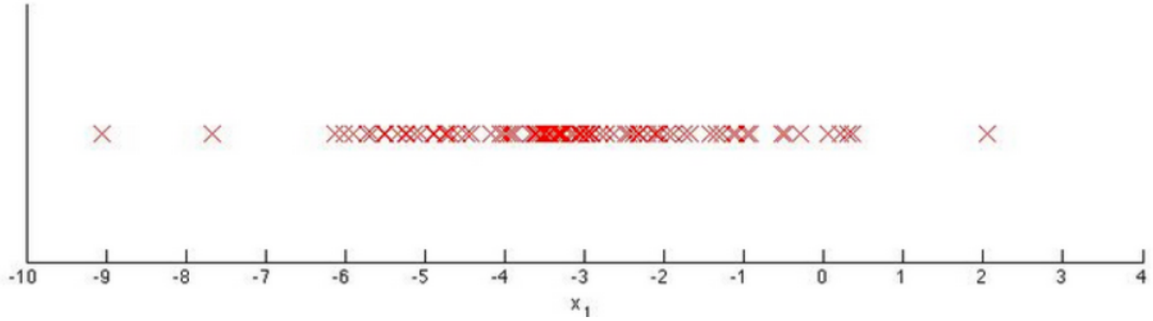
   We need a feature that takes into consideration both variables, since one changes and one doesn't. Multiplying the two factors don't work, as two regular values for x_1 and x_2 multiplied together could net the same value for their product as a large x_1 and small x_2. We will want the ratio between them, so finding a large ratio will identify this anomaly.

   $x_3 = \dfrac{x_1}{x_2}$

4. Which of the following are true? Check all that apply.

   (a) When developing an anomaly detection system, it is often useful to select an appropriate numerical performance metric to evaluate the effectiveness of the learning algorithm.

5. You have a 1-D dataset $\{x^{(1)}, ..., x^{(m)}\}$ and you want to detect outliers in the dataset. You first plot outliers in the dataset and it loos like this:



Suppose you fit the Guassian distribution parameters μ_1 and σ_1^2 to this dataset. Which of the following values for μ_1 and σ_1^2 might you get?

The points are concentrated around -3, which will be the mean. Most, if not all the points will be housed within 2 standard deviations of the mean. That seems to be around -7 and 1 respectively, so 1 standard deviation is 2. Variance is the standard deviation squared, so it would be 4. Thus: $\mu_1 = -3, \sigma_1^2 = 4$