anomaly detection system construction

developing and evaluating an anomaly detection system

the importance of real · number evaluation

when developing learning algorithm (choosing features, etc…), making decisions is easier if a method of algorithm evaluation is available

assuming procession of labeled data, of anomalous and non-anomalous examples ($y = 0$ if normal, $y = 1$ if anomalous)

training set: $\left(x^{(1)}, x^{(2)}, \dots, x^{(m)}\right)$ assuming normal examples/non-anomalous
cross validation set: $\left(x_{cv}^{(1)}, y_{cv}^{(1)}, \dots, x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})}\right)$
test set: $\left(x_{test}^{(1)}, y_{test}^{(1)}, \dots, x_{test}^{(m_{test})}, y_{test}^{(m_{test})}\right)$

motivating example

10,000 good (normal) examples
20 flawed (anomalous) examples

$\mu_1, \sigma_1^2, \dots, \mu_n, \sigma_n^2$

training set: $\boxed{6000}$ good examples ($y = 0$) $\quad p(x) = p(x_1; \mu_1, \sigma_1^2), \dots, p(x_n; \mu_n, \sigma_n^2)$
cross validation: 2000 good examples ($y = 0$), 10 anomalous ($y = 1$)
test set: 2000 good examples ($y = 0$), 10 anomalous ($y = 1$)

alternative (**not recommended**, same examples in cv and test):

training set: 6000 good examples
cross validation: 4000 good examples ($y = 0$), 10 anomalous ($y = 1$)
test set: 4000 good examples ($y = 0$), 10 anomalous ($y = 1$)

algorithm evaluation

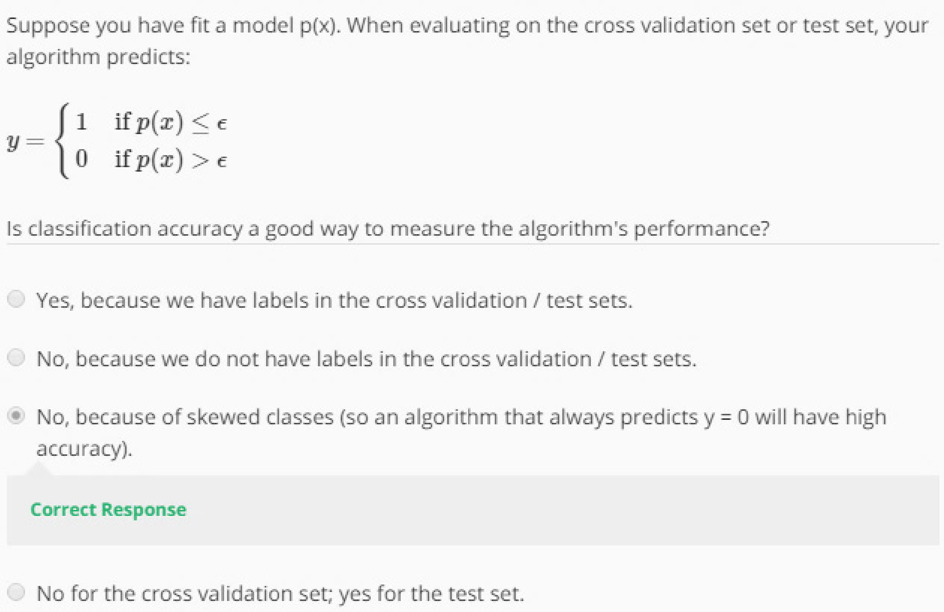
fit model $p(x)$ on the training set $\{x^{(1)}, \dots, x^{(m)}\}$
predict x on the cross validation/test example

$$y = \begin{cases} 1 & \text{if } p(x) < \varepsilon \text{ (anomoly)} \\ 0 & \text{if } p(x) \geq \varepsilon \text{ (normal)} \end{cases}$$

possible evaluation metrics: $\left(x_{test}^{(i)}, y_{test}^{(i)}\right)$

- true positive, false positive, false negative, true negative
- precision/recall
- F_1-score

can also use cross validation set to choose parameter ε



Suppose you have fit a model p(x). When evaluating on the cross validation set or test set, your algorithm predicts:

$$y = \begin{cases} 1 & \text{if } p(x) \leq \epsilon \\ 0 & \text{if } p(x) > \epsilon \end{cases}$$

Is classification accuracy a good way to measure the algorithm's performance?

○ Yes, because we have labels in the cross validation / test sets.

○ No, because we do not have labels in the cross validation / test sets.

● No, because of skewed classes (so an algorithm that always predicts y = 0 will have high accuracy).

Correct Response

○ No for the cross validation set; yes for the test set.

anomaly detection versus supervised learning

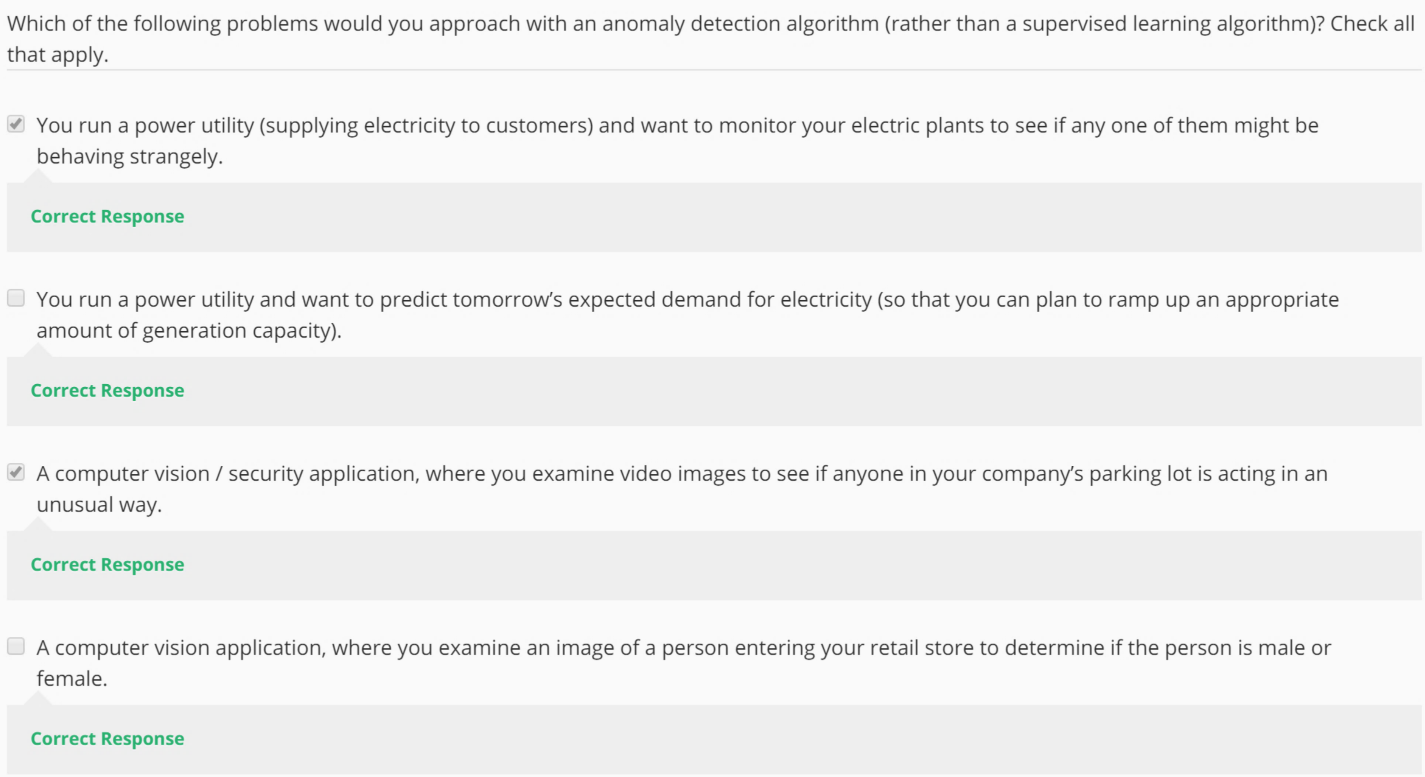
when processing labeled data, the question arises as to using a supervised learning method for a portion of the data (logistic regression, etc...) as opposed to an anomaly detection algorithm. the following example compares the advantages of each:

anomaly detection	supervised learning
very small number of positive examples $(y = 1)$ (0-20 is common)	large number of positive and negative examples
large number of negative $(y = 0)$ examples $p(x)$	
many different "types" of anomalies. difficult for any algorithm to learn from positive examples what the anomalies might appear as;	enough positive examples for an algorithm to obtain a sense of what the positive examples will appear as; future positive examples are likely to be similar to those in the training set
future anomalies may appear nothing like any of the preceding anomalous examples witnessed.	

application examples

anomaly detection: fraud detection, manufacturing, monitoring data centers

supervised learning: email spam classification, weather prediction, cancer research



Which of the following problems would you approach with an anomaly detection algorithm (rather than a supervised learning algorithm)? Check all that apply.

☑ You run a power utility (supplying electricity to customers) and want to monitor your electric plants to see if any one of them might be behaving strangely.

Correct Response

□ You run a power utility and want to predict tomorrow's expected demand for electricity (so that you can plan to ramp up an appropriate amount of generation capacity).

Correct Response

☑ A computer vision / security application, where you examine video images to see if anyone in your company's parking lot is acting in an unusual way.

Correct Response

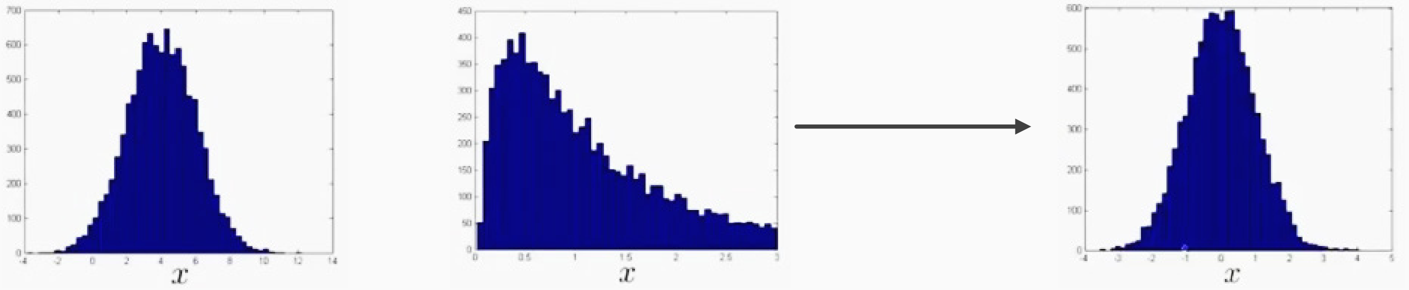
□ A computer vision application, where you examine an image of a person entering your retail store to determine if the person is male or female.

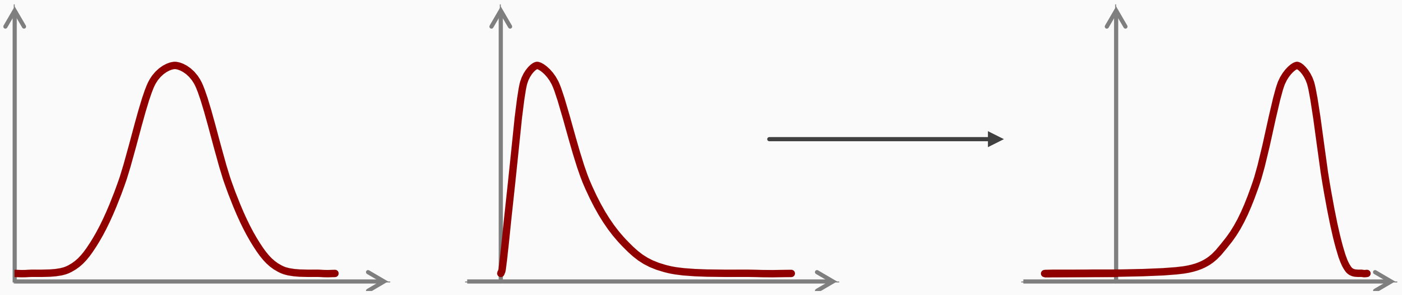
Correct Response

choosing what features to use

transformations of non · gaussian features

$p(x_1; \mu_1, \sigma_1^2)$





normal distribution $x_1 \leftarrow \log(x_1)$ $x_2 \leftarrow \log(x_2 + c)$ $x_3 \leftarrow \sqrt{x_3} = x_3^{\frac{1}{2}}$ $x_4 \leftarrow \sqrt[3]{x_4} = x_4^{\frac{1}{3}}$

transformations to the dataset can be applied with various methods to achieve a gaussian (normal) distribution from that of a skewed dataset

error analysis for anomaly detection

creating features for an anomaly detection problem by observing misclassifications on cross validation dataset

expectation of $p(x)$ as large for normal examples x

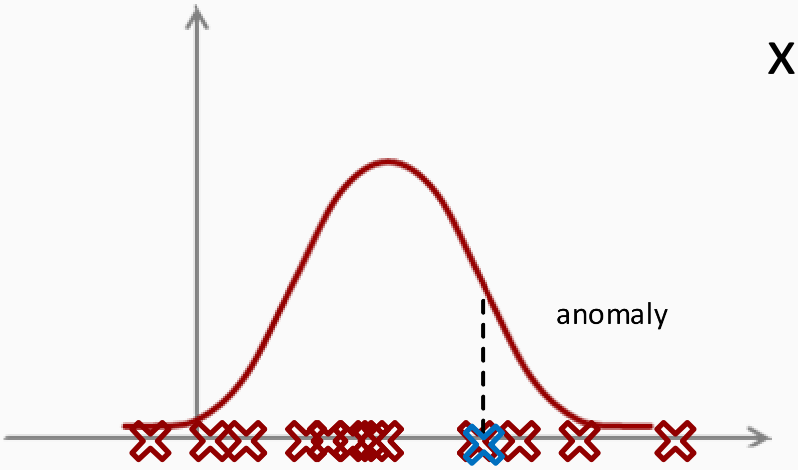
$p(x)$ as small for anomalous examples x

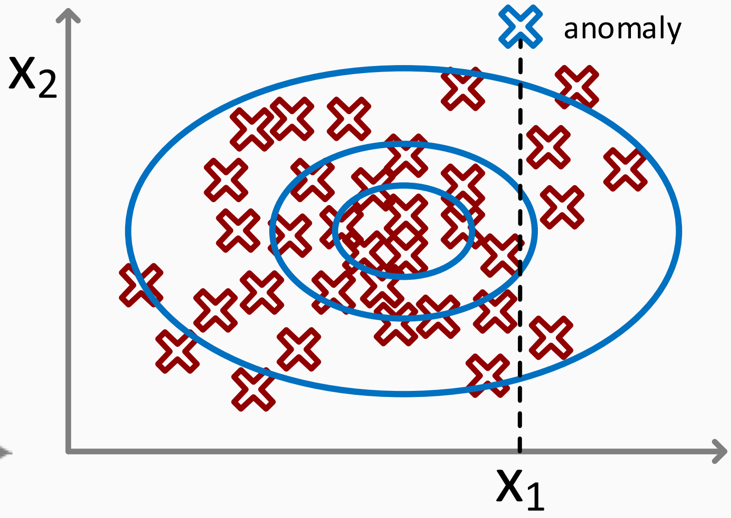
the common problem: $p(x)$ is comparable (e.g., both are large)
for normal and anomalous examples

after fitting a Gaussian distribution to a dataset, a new example known to be anomalous is misclassified as normal according to the current algorithm's performance

a new feature (x_2) is created and takes on an unusual value when plotted against the same new example known to be anomalous

the process involves physical interactions and observation of the data as it is being introduced into the model in order to create new features for anomaly detection





ideology for choosing features for anomaly detection

choose features which potentially take on unusually large or small values in the event of an anomaly

x_1 = measured feature

x_2 = measured feature

x_3 = measured feature

x_4 = measured feature

assuming two features have a particular relationship, it might benefit anomaly detection to relate them against a new feature:

$\frac{x_1}{x_2} = x_5$ or $\frac{(x_4)^2}{x_3} = x_6$ etc…