ML w9 problem motivation

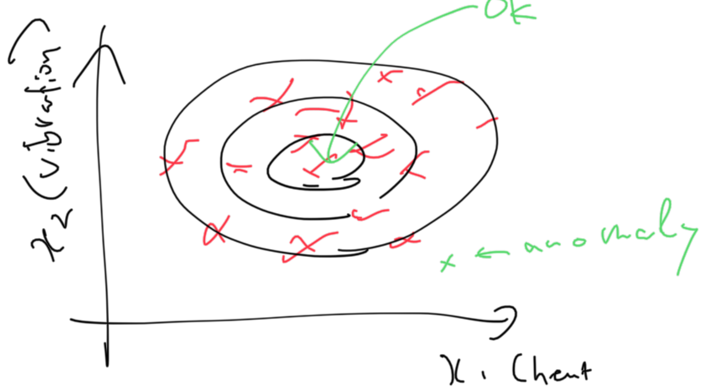
Problem Motivation

Density estimation

ϵ epsilon

$p(X_{test}) < \epsilon \rightarrow$ flag anomaly

$p(X_{test}) \geq \epsilon \rightarrow$ ok



Anomaly detection example

→ Fraud detection
$x^{(i)}$ features

→ Manufacturing

→ Monitoring

Gausian Distribution

$x \in \mathbb{R}$ (x is a real number)

Gausian with mean μ, variance σ^2

$x \sim N(\mu, \sigma^2)$

σ = standard deviation

σ^2 = variance

↑ normal
distributed as

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

0 / 1点

- [x] In a computer chip fabrication plant, identify microchips that might be defective.

正解
The defective chips are the anomalies you are looking for by modeling the properties of non-defective chips.

- [x] Given data from credit card transactions, classify each transaction according to type of purchase (for example: food, transportation, clothing).

これを選択しないでください
Anomaly detection is not appropriate for a traditional classification problem.

- [ ] From a large set of primary care patient records, identify individuals who might have unusual health conditions.
- [ ] From a large set of hospital patient records, predict which patients have a particular disease (say, the flu).

2. Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when $p(x)$ is less than ε, and you find on the cross-validation set that it is missing many fradulent transactions (i.e., failing to flag them as anomalies). What should you do?

1 / 1点

- ( ) Decrease ε
- (•) Increase ε

正解
By increasing ε, you will flag more anomalies, as desired.

3. Suppose you are developing an anomaly detection system to catch manufacturing defects in airplane engines. You model uses

1 / 1点

$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$.

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 ranges of values. What additional feature x_3 should you create to capture these types of anomalies:

- ( ) $x_3 = x_1 \times x_2^2$
- ( ) $x_3 = x_1^2 \times x_2^2$
- ( ) $x_3 = (x_1 + x_2)^2$
- (•) $x_3 = \frac{x_1}{x_2}$

正解
This is correct, as it will take on large values for anomalous examples and smaller values for normal examples.

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

1 / 1点

- [x] If you do not have any labeled data (or if all your data has label $y = 0$), then is is still possible to learn $p(x)$, but it may be harder to evaluate the system or choose a good value of ϵ.

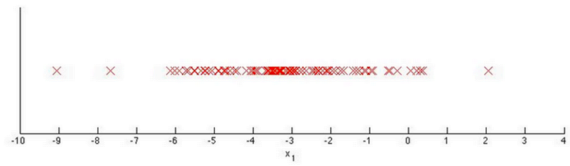
正解
Only negative examples are used in training, but it is good to have some labeled data of both types for cross-validation.

- [ ] If you have a large labeled training set with many positive examples and many negative examples, the anomaly detection algorithm will likely perform just as well as a supervised learning algorithm such as an SVM.
- [ ] If you are developing an anomaly detection system, there is no way to make use of labeled data to improve your system.
- [x] When choosing features for an anomaly detection system, it is a good idea to look for features that take on unusually large or small values for (mainly the) anomalous examples.

正解
These are good features, as they will lie outside the learned model, so you will have small values for $p(x)$ with these examples.

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

1 / 1点



Suppose you fit the gaussian 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?

- (●) $\mu_1 = -3, \sigma_1^2 = 4$
- ( ) $\mu_1 = -6, \sigma_1^2 = 4$
- ( ) $\mu_1 = -3, \sigma_1^2 = 2$
- ( ) $\mu_1 = -6, \sigma_1^2 = 2$

正解
This is correct, as the data are centered around -3 and tail most of the points lie in [-5, -1].

Predicting Movie Rate.

j

i	user 1	2	3	4
movie A	?	?	4	5
B	1	3	4	5
C	?	5	?	4
D	1	?	2	3

n_u = no. users = 4

n_m = no. movies = 4

$r(i,j) = 1$ if user j has rated movie i

1. Suppose you run a bookstore, and have ratings (1 to 5 stars)

of books. Your collaborative filtering algorithm has learned

a parameter vector $\theta^{(j)}$ for user j, and a feature

vector $x^{(i)}$ for each book. You would like to compute the

"training error", meaning the average squared error of your

system's predictions on all the ratings that you have gotten

from your users. Which of these are correct ways of doing so (check all that apply)?

For this problem, let m be the total number of ratings you

have gotten from your users. (Another way of saying this is

that $m = \sum_{i=1}^{n_m} \sum_{j=1}^{n_u} r(i,j)$). [Hint: Two of the four options below are correct.]

- [x] $\frac{1}{m} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} (\sum_{k=1}^{n} (\theta^{(j)})_k x_k^{(i)} - y^{(i,j)})^2$
- [ ] $\frac{1}{m} \sum_{(i,j):r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - r(i,j))^2$
- [x] $\frac{1}{m} \sum_{(i,j):r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2$
- [ ] $\frac{1}{m} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} (\sum_{k=1}^{n} (\theta^{(k)})_j x_i^{(k)} - y^{(i,j)})^2$

2. In which of the following situations will a collaborative filtering system be the most appropriate learning algorithm (compared to linear or logistic regression)?

- [ ] You're an artist and hand-paint portraits for your clients. Each client gets a different portrait (of themselves) and gives you 1-5 star rating feedback, and each client purchases at most 1 portrait. You'd like to predict what rating your next customer will give you.
- [x] You run an online bookstore and collect the ratings of many users. You want to use this to identify what books are "similar" to each other (i.e., if one user likes a certain book, what are other books that she might also like?)
- [x] You own a clothing store that sells many styles and brands of jeans. You have collected reviews of the different styles and brands from frequent shoppers, and you want to use these reviews to offer those shoppers discounts on the jeans you think they are most likely to purchase
- [ ] You manage an online bookstore and you have the book ratings from many users. You want to learn to predict the expected sales volume (number of books sold) as a function of the average rating of a book.

3. You run a movie empire, and want to build a movie recommendation system based on collaborative filtering. There were three popular review websites (which we'll call A, B and C) which users to go to rate movies, and you have just acquired all three companies that run these websites. You'd like to merge the three companies' datasets together to build a single/unified system. On website A, users rank a movie as having 1 through 5 stars. On website B, users rank on a scale of 1 - 10, and decimal values (e.g., 7.5) are allowed. On website C, the ratings are from 1 to 100. You also have enough information to identify users/movies on one website with users/movies on a different website. Which of the following statements is true?

- (●) You can merge the three datasets into one, but you should first normalize each dataset's ratings (say rescale each dataset's ratings to a 0-1 range).
- ( ) You can combine all three training sets into one as long as your perform mean normalization and feature scaling **after** you merge the data.
- ( ) Assuming that there is at least one movie/user in one database that doesn't also appear in a second database, there is no sound way to merge the datasets, because of the missing data.
- ( ) It is not possible to combine these websites' data. You must build three separate recommendation systems.

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

- [ ] For collaborative filtering, the optimization algorithm you should use is gradient descent. In particular, you cannot use more advanced optimization algorithms (L-BFGS/conjugate gradient/etc.) for collaborative filtering, since you have to solve for both the $x^{(i)}$'s and $\theta^{(j)}$'s simultaneously.

- [x] For collaborative filtering, it is possible to use one of the advanced optimization algoirthms (L-BFGS/conjugate gradient/etc.) to solve for both the $x^{(i)}$'s and $\theta^{(j)}$'s simultaneously.

- [ ] Suppose you are writing a recommender system to predict a user's book preferences. In order to build such a system, you need that user to rate all the other books in your training set.

- [x] Even if each user has rated only a small fraction of all of your products (so $r(i,j) = 0$ for the vast majority of (i,j) pairs), you can still build a recommender system by using collaborative filtering.

5. Suppose you have two matrices A and B, where A is 5x3 and B is 3x5. Their product is $C = AB$, a 5x5 matrix. Furthermore, you have a 5x5 matrix R where every entry is 0 or 1. You want to find the sum of all elements $C(i,j)$ for which the corresponding $R(i,j)$ is 1, and ignore all elements $C(i,j)$ where $R(i,j) = 0$. One way to do so is the following code:

```
C = A * B;
total = 0;
for i = 1:5
  for j = 1:5
    if (R(i,j) == 1)
      total = total + C(i,j);
    end
  end
end
```

Which of the following pieces of Octave code will also correctly compute this total? Check all that apply. Assume all options are in code.

- [x] total = sum(sum((A * B) .* R))

- [ ] C = A * B; total = sum(sum(C(R == 1)));

- [x] C = (A * B) * R; total = sum(C(:));

- [ ] total = sum(sum(A(R == 1) * B(R == 1));

5. Suppose you have two matrices A and B, where A is 5x3 and B is 3x5. Their product is $C = AB$, a 5x5 matrix. Furthermore, you have a 5x5 matrix R where every entry is 0 or 1. You want to find the sum of all elements $C(i,j)$ for which the corresponding $R(i,j)$ is 1, and ignore all elements $C(i,j)$ where $R(i,j) = 0$. One way to do so is the following code:

```
C = A * B;
total = 0;
for i = 1:5
  for j = 1:5
    if (R(i,j) == 1)
      total = total + C(i,j);
    end
  end
end
```

Which of the following pieces of Octave code will also correctly compute this total? Check all that apply. Assume all options are in code.

0 / 1点

⊗ 不正解