Week 9

1

Anormaly Detection

Dataset : $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ 手本 (normal examples)

x_{test} ← $P(x_{test}) < \varepsilon$ → anomaly

$P(x_{test}) \geq \varepsilon$ → OK

Fraud detection

$x^{(i)}$ = features of user i's activities $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)}\}$

Model $p(x)$ from data.

Identify unusual users by checking which have $p(x) < \varepsilon$

[講義中の Quiz]

異常検知システムは $p(x) \leq \varepsilon$ のとき x のフラグを立てる。

システムがフラグを立てやすい時、どうする? ← false positive (教師あり学習での呼び方)

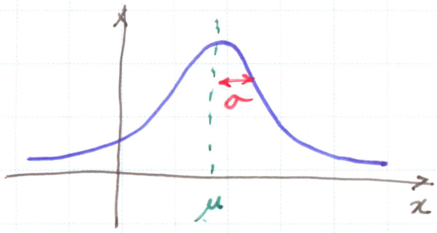
$p(x) \leq \varepsilon$ をみたす x を減らしたい → ε を小さく

Week 9　　2

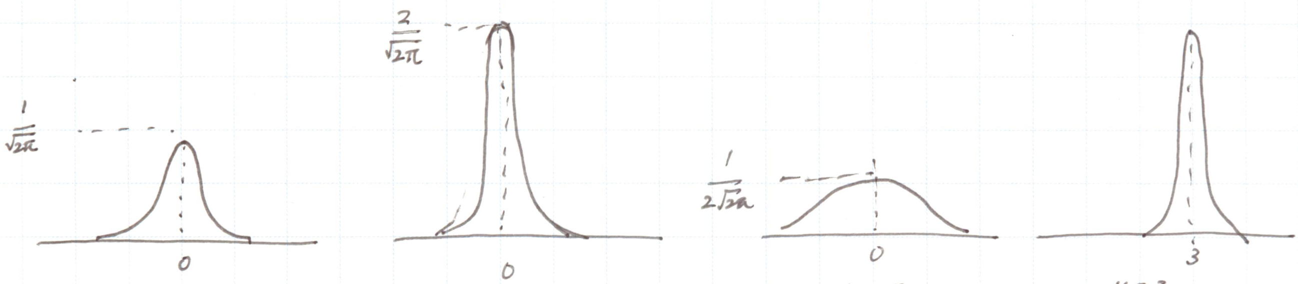
Gaussian Distribution (= Normal Distribution)

正規分布

$x \in \mathbb{R}$, if x is a distributed Gaussian with mean μ, variance σ^2



$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$\mu=0$, $\sigma=1$　　$\mu=0$, $\sigma=0.5$　　$\mu=0$, $\sigma=2$　　$\mu=3$, $\sigma=0.5$

Parameter estimation

Dataset : $\{x^{(1)}, \dots, x^{(m)}\}$,　$x^{(i)} \in \mathbb{R}$　　$x \sim N(\mu, \sigma^2)$

確率変数xが1次元の場合　　と表記する

$x^{(i)} \in \mathbb{R}^n$　　$x \sim N_n(\mu, \Sigma)$

n次元の場合

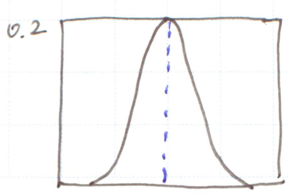
与えられたdatasetに対して μやσ^2の値を推定したい

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}, \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)^2$$

統計では$(m-1)$を使うが 機械学習では mを使う

[講義中のQuiz]

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$-3 = \mu$

$$\frac{1}{\sqrt{2\pi}\,\sigma} = 0.2 = \frac{1}{5}$$

$$6.28\overline{)25.00}\;\; 4, \quad 25.12, \quad -.12$$

$$\therefore \sqrt{2\pi}\,\sigma = 5$$

$$\sigma = \sqrt{\frac{25}{2\pi}} \simeq \sqrt{4} = 2$$

$$\therefore p(x) = \frac{1}{\sqrt{2\pi}\,\underset{\sigma}{2}}\exp\left(-\frac{(x\overset{-\mu}{+3})^2}{2\times \underset{\sigma^2}{4}}\right)$$

Week 9　　3

Density estimation　（確率密度関数の推定）

Training set : $\{x^{(1)}, \ldots, x^{(m)}\}$
$x \in \mathbb{R}^n$

$$\begin{cases} x_1 \sim N(\mu_1, \sigma_1^2) \\ \vdots \\ x_n \sim N(\mu_n, \sigma_n^2) \end{cases}$$

$$P(x) = P(x_1; \mu_1, \sigma_1^2)\, P(x_2; \mu_2, \sigma_2^2) \cdots P(x_n; \mu_n, \sigma_n^2)$$

$$= \prod_{j=1}^{n} P(x_j; \mu_j, \sigma_j^2) = \underbrace{\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)}$$

Anomaly if $P(x) \leq \varepsilon$

[~~講義~~中の Quiz]

$\{x^{(1)}, \ldots, x^{(m)}\}$,　$\mu_j \in \mathbb{R}$,　$\sigma_j^2 \in \mathbb{R}$

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}, \quad \sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2$$

Week 9 　　4

Building an Anomaly Detection System

real-number evaluation は重要

labeled data （教師ありデータ）

$$\begin{cases} y=0 & \text{if normal} \\ y=1 & \text{if anormal} \end{cases}$$

Training set $\{x^{(1)}, \ldots, x^{(m)}\}$ ← 全て normal なデータ
ラベル付けなし

Cross validation set $(x_{cv}^{(1)}, y_{cv}^{(1)}), \ldots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$ ← anormal なデータも含む

Test set $(x_{test}^{(1)}, y_{test}^{(1)}), \ldots, (x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$

航空機のエンジンの例

10000 ← good engine (normal)
20 ← flawed engine (anormal)

Training set : 6000 good
CV set : 2000 good $(y=0)$, 10 anormal $(y=1)$
Test set : 2000 good $(y=0)$, 10 anormal $(y=1)$

Algorithm evaluation

モデル $p(x)$ を training set $\{x^{(1)}, \ldots, x^{(m)}\}$ に fit させる。

cross validation set に対して

$$y = \begin{cases} 1 & \text{if } p(x) \le \varepsilon \quad \text{(anormal)} \\ 0 & \text{if } p(x) > \varepsilon \quad \text{(normal)} \end{cases}$$

評価基準

- True Positive, false positive, true negative
- Precision / Recall
- F1 - score

ε パラメータを決定するのに CV set を使う。

		実際 1	実際 0
予測	1	true positive	false positive
予測	0	false negative	true negative

Week 9　　5

注意
accuracyは特定の計算方法を指している

$$\frac{(1)+(4)}{(1)+(2)+(3)+(4)}$$

		actual 1	actual 0
予測	1	(1)	(2)
	0	(3)	(4)

[講義中のQuiz]

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

アルゴリズムの性能を評価するのにclassification accurayはよい方法か？

① yes. cvとtest setにはlabelがついている（答，教師）　×

② No. cvとtestにはlabelがない　×

③ No. skewed classなのでいつも y=0 が良い結果となる。　○

④ No. for the cv. yes for test.　×

Anormally Detection vs. Supervised Learning

Anormally detection	vs. Supervised learning
positive事例 (y=1) がとても少ない (0〜20個)	positive事例もnegative事例もたくさんある。
negative事例 (y=0) はたくさんある。	positiveを学習するのに十分なだけのpositive例が与えられる。
anomalityはいろいろなtypeがあるので、positive事例からはanomalyがどのようなものか、学習するのは難しい。	未来に出会うpositive例も、まぁ、過去のpositiveと似ているだろう。
将来出会うanomalyは、今まで出会ったanomaliesとは全く異っているかもしれない	
Fraud detection Manufacturing Monitoring machines in data center	Email spam classification weather prediction (sunny/rainy/etc) Cancer classification

日本総研
The Japan Research Institute, Limited

Week 9

[~~講義~~中のQuiz]

anomaly detection を用いる（supervised learningではなく）べき問題を選べ

① 電源設備を動作させていて、動作がおかしいかモニターする ○

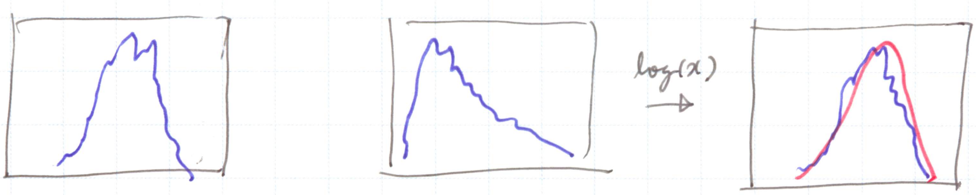
② 〃 明日の必要電力を予測する ×

④ computer vision/securityアプリケーションで、店に入ってきた人が男か女かイメージから予測する ×

③ 〃 、店の駐車場が普通ではない使い方をされているかvideoから予測する ○

Choosing what features to use

Non-Gaussian features



$x_1 \leftarrow \log(x_1)$

$x_2 \leftarrow \log(x_2 + C)$

$x_3 \leftarrow \sqrt{x_3} = (x_3)^{\frac{1}{2}}$

$x_4 \leftarrow (x_4)^{\frac{1}{3}}$

ガウス分布（正規分布）に似せる.

Error analysis for anomaly detection

Want $p(x)$ { large for normal example x.
small for anomalous 〃

Most Common Problem

$p(x)$ は comparable { normal example
anomalous 〃

[講義中の Quiz]

anomaly detection → 性能悪い
normal example に多し $p(x)$ 大
anomalies in CV に多し $p(x)$ 大

どう変更するよいか?

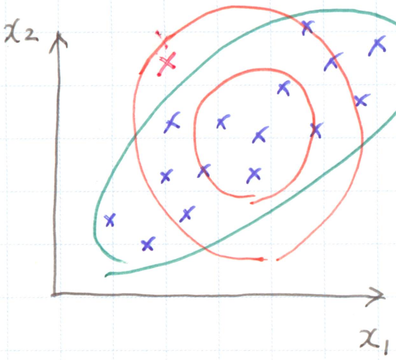
① 少い featureで-　　✗

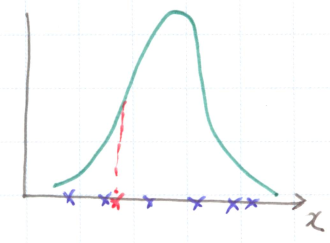
② normal example と anormal example を区別するために もっと多くの feature をいれる ✗ Oが正解

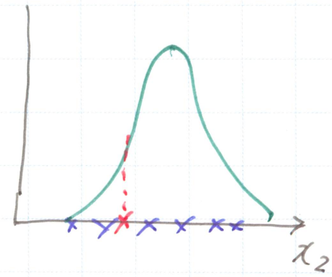
③ 大量の (normal example の) training set を入手する　　✗

④ ε をかえる　　✗ ε を大きくすべき

Multivariate Gaussian Distribution (Optional)







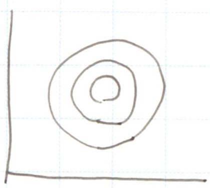
$x_1, x_2, \cdots$ それぞれで正規分布を判断しては上の × は anomaly detection できない

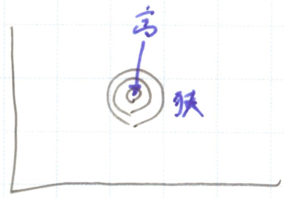
$x \in \mathbb{R}^n$. $p(x_1), p(x_2), \ldots$ を個別に model化してはいけない。
全ての x について一度にモデル化 $p(x)$ すべきだ。

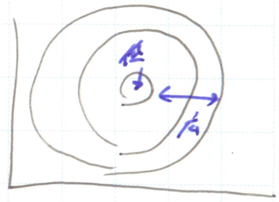
パラメーター： $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$ (covariance matrix)
共分散行列

$$f(x) = \frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

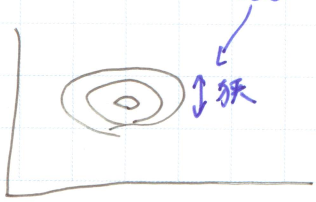
$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6 \end{pmatrix}$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$

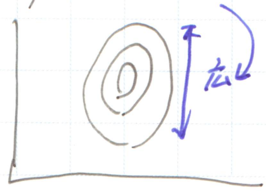






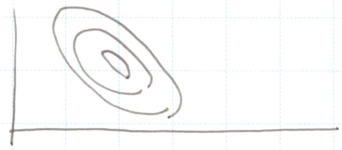
$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.6 \end{pmatrix}$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$



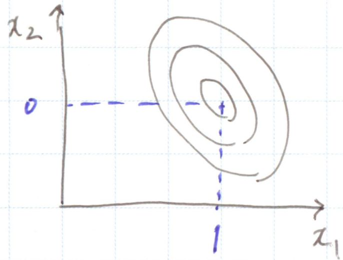


Week 9　　9

$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$,　$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 0 \end{pmatrix}$



[講義中のQuiz]



$\mu = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$

Anomaly Detection using the Multivariate Gaussian Distribution

多変量正規分布

パラメータ　μ, Σ

$$p(x; \mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

パラメータ fitting

training set　$\{x^{(1)}, \dots, x^{(m)}\}$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}, \quad \Sigma = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

これで得られた μ と Σ を anomaly detection に利用する.

Week 9 10

Anomaly detection with the multivariate Gaussian

1. model $p(x)$ を fit する

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

2. 新しい example x が与えられるので $p(x)$ を計算する

$$p(x) = \frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

if $p(x) < \varepsilon$, フラグを立てる.

Relationship to original model

Original model : $p(x) = p(x_1; \mu_1, \sigma_1^2) \cdot p(x_2; \mu_2, \sigma_2^2) \cdot \cdots \cdot p(x_n; \mu_n, \sigma_n^2)$

Corresponds to multivariate Gaussian

$$p(x; \mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where $\Sigma = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix}$

- feature の間の相関を自動的に扱うことができる
- $m > n$ が必須 ($m > 10n$ が理想)
  そうでないと Σ の逆行列が存在しない.

- feature 間の相関を扱うためには $x_3 = \frac{x_2}{x_1}$ のように自分で考えて相関をうまく表す feature を作って, 追加する必要がある
- 計算量は小さい ($n = 10000$ 以上でもOK)
- m (training set のサイズ) が小さくても ok.

Σ が singular なのは (特異行列, 非可逆) $\begin{cases} m < n \text{ の時} \\ \text{or} \\ \text{feature が冗長の場合 (線形に従属している)} \end{cases}$

Week 9　　11

[~~講義~~中の Quiz]

$\{x^{(1)}, \dots, x^{(m)}\}$ where $x^{(i)} \in \mathbb{R}^n$. anomaly detection を適用する。

① original model $\prod_{i=1}^{n} p(x_i; \mu_i, \sigma_i^2)$ は $p(x; \mu, \Sigma)$ の輪郭が軸 aligned ならば multivariate Gaussian に対応している。 ~~×~~ ○

teature の間に相関が一切なければ

軸に沿っているということは相関が無いということ

② multivariate Gaussian は $m < n$ の時 有利である ×

Σが非可逆になるのでダメ

③ multivariate Gaussian は feature 間の相関を自動的に扱ってくれる ○

④ original model は 計算量的に少なくて済む。 n が大きい時に有利 ○

日本総研 The Japan Research Institute, Limited

Predicting Movie Rating

Problem Formulation

Recommendation System … 機械学習のよい応用例
featureを自動で選択する例になる。

例 movie rating

movie \ user	user1	2	3	4
movie 1	5	5	0	0
2	5	?	?	0
3	2	4	0	?
4	0	0	5	4
5	0	0	5	?

← 星の数

n_u : ユーザ数

n_m : 映画の数

$r(i,j) = 1$ if ユーザjが映画iをratingしている

$y^{(i,j)}$ = rating (ユーザj, 映画i)

[講義中のQuiz]

	user 1	2	3
Movie 1	0	1	?
2	$\boxed{?}$	5	5

movie user

$r(2,1) = 0$

$y^{(2,1)} = ?$

Week 9　　13

Content Based Recommendations

例

movie \ user	user 1	2	3	4	romance x_1	action x_2
movie 1	5	5	0	0	0.9	0
2	5	?	?	0	1.0	0.01
3	?	4	0	?	0.99	0
4	0	0	5	4	0.1	1.0
5	0	0	5	?	0	0.9

$x^{(1)} = \begin{pmatrix} 1 \\ 0.9 \\ 0 \end{pmatrix}$　Bias?

$n_u = 4$, $n_m = 5$　　$n = 2$　+1

各ユーザ j に対して，パラメータ $\theta^{(j)} \in \mathbb{R}^3$ を学習させる.

ユーザ j について，映画 i の rating を $(\theta^{(j)})^T x^{(i)}$ と予測する.

[使用例]　$x^{(3)} = \begin{pmatrix} 1 \\ 0.99 \\ 0 \end{pmatrix}$,　ここで $\theta^{(1)} = \begin{pmatrix} 0 \\ 5 \\ 0 \end{pmatrix}$ と学習したのであれば

$$y^{(1,3)} = \theta^{(1)T} x^{(3)} = (0 \;\; 5 \;\; 0) \begin{pmatrix} 1 \\ 0.99 \\ 0 \end{pmatrix} = 4.95$$

[講義中の Quiz]

	user 1	2	3	4	romance x_1	action x_2
Movie 1	5	5	0	0	0.9	0
2	5	?	?	0	1.0	0.01
3	?	4	0	?	0.99	0
4	0	0	5	4	0.1	1.0
5	0	0	5	?	0	0.9

これの rating を predict する $\theta^{(j)}$

$i = 1, 2$　$x^{(i)} = \begin{pmatrix} 1 \\ 大 \\ 小 \end{pmatrix}$

$i = 3, 4$　$x^{(i)} = \begin{pmatrix} 1 \\ 小 \\ 大 \end{pmatrix}$

$x_0 = 1$。$\theta^{(3)}$ として適切なのはどれ

① $\theta^{(3)} = \begin{pmatrix} 0 \\ 5 \\ 0 \end{pmatrix}$,　② $\theta^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$　③ $\theta^{(3)} = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix}$,　④ $\theta^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix}$

∵ ユーザ3はユーザ4と同じクラスで，これを高く評価するのは x_2.

$\theta^{(i)T} x^{(j)}$ で x_2 の係数は $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ ← コレ

$x_2 = 1$ のとき，値として5を出力したいので ④ が正解.

Problem Formulation

$r(i,j) = \begin{cases} 1 & \text{if } ユーザjが映画iについてratingしていたら \\ 0 & \text{otherwise} \end{cases}$

$y^{(i,j)}$ = ユーザjが映画iに与えたratingの値

$\theta^{(j)}$ = ユーザjについてのパラメータ・ベクタ

$x^{(i)}$ = 映画iに対するfeatureベクタ

ユーザjが映画iについて与えるratingの予測値 $= (\theta^{(j)})^T x^{(i)}$

$m^{(j)}$ = ユーザjがratingを与えた映画の数

$\theta^{(j)}$を学習する. $\theta^{(j)} \in \mathbb{R}^{n+1}$ Regression(回帰)で解ける

$$\min_{\theta^{(j)}} \frac{1}{2m^{(j)}} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2m^{(j)}} \sum_{k=1}^{n} (\theta_k^{(j)})^2$$

コレは無してもよい. 定数倍だから

$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n_u)}$を学習する

$$\min_{\theta^{(1)},\ldots,\theta^{(n_u)}} \underbrace{\frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^{n} (\theta_k^{(j)})^2}_{J(\theta^{(1)},\ldots,\theta^{(n_u)})}$$

Gradient descent update

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right) x_k^{(i)} \quad (k=0のとき)$$

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \left(\sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right) x_k^{(i)} + \lambda \theta_k^{(j)} \right) \quad (k \neq 0のとき)$$

Week 9 15

Collaborative Filtering

[講義中の Quiz]

	user 1	2	3	romance x_1
Movie 1	0	1.5	2.5	? ← 0.5

$x = \begin{pmatrix} 1 \\ x_1 \end{pmatrix}$

$\theta^{(1)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \theta^{(2)} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \ \theta^{(3)} = \begin{pmatrix} 0 \\ 5 \end{pmatrix}$

$y^{(1,3)} = (0\ 5)\begin{pmatrix} 1 \\ x_2 \end{pmatrix} = 5x_1 = 2.5$

$y^{(2,3)} = (0\ 3)\begin{pmatrix} 1 \\ x_2 \end{pmatrix} = 3x_1 = 1.5$

$\therefore x_1 = 0.5$ (答)

Optimization algorithm

$\theta^{(1)}, \dots, \theta^{(n_u)}$ が与えられた時、$x^{(i)}$ を学習する

$$\min_{x^{(i)}} \frac{1}{2} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{k=1}^{n} (x_k^{(i)})^2$$

$\theta^{(1)}, \dots \theta^{(n_u)}$ が与えられた時、$x^{(1)}, x^{(2)}, \dots, x^{(n_m)}$ を学習する

$$\min_{x^{(1)}, \dots x^{(n_m)}} \frac{1}{2} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{1}{2} \sum_{i=1}^{n_m} \sum_{k=1}^{n} (x_k^{(i)})^2$$

Collaborative filtering

[講義中の Quiz]

$$\min_{x^{(1)}, \dots, x^{(n_m)}} \frac{1}{2} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^{n} (x_k^{(i)})^2$$

$i \neq 0$ に対する gradient descent の正しい更新は？

$$x_k^{(i)} := x_k^{(i)} \underset{\sim}{-} \alpha \left(\sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) \theta_k^{(j)} + \underline{\lambda x_k^{(i)}} \right)$$

←

$i \neq 0$ だから必要

Week 9　　16

Collaborative filtering （協調的フィルタリング）

$x^{(1)}, x^{(2)}, \dots, x^{(n_m)} \longrightarrow \theta^{(1)}, \dots \theta^{(n_u)}$ を評価する

$\theta^{(1)}, \dots, \theta^{(n_u)} \longrightarrow x^{(1)}, \dots, x^{(n_m)}$ を評価する

Collaborative filtering algorithm

$$\min_{\theta^{(1)},\dots,\theta^{(n_u)}} \frac{1}{2}\sum_{j=1}^{n_u}\sum_{i:r(i,j)=1}\left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2$$

$$\min_{x^{(1)},\dots,x^{(n_m)}} \frac{1}{2}\sum_{i=1}^{n_m}\sum_{j:r(i,j)=1}\left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{n_m}\sum_{k=1}^{n}(x_k^{(i)})^2$$

θ と x を同時に解くことができる

Minimizing $x^{(1)}, \dots, x^{(n_m)}$ and $\theta^{(1)}, \dots, \theta^{(n_u)}$ simultaneously

$$J(x^{(1)},\dots,x^{(n_m)},\theta^{(1)},\dots,\theta^{(n_u)}) = \frac{1}{2}\sum_{(i,j):r(i,j)=1}\left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{n_m}\sum_{k=1}^{n}(x_k^{(i)})^2 + \frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2$$

と $J(\cdots)$ を定義して

$$\min_{\substack{x^{(1)},\dots,x^{(n_m)}\\ \theta^{(1)},\dots,\theta^{(n_u)}}} J(x^{(1)},\dots,x^{(n_m)},\theta^{(1)},\dots,\theta^{(n_u)})$$

もしも常に1となるfeatureが必要であれば、それを学習してfeature n 選んでくれるはずだ

このとき feature を学習することになるので バイアス項 x_0 はいらない。

$\therefore x \in \mathbb{R}^n$

$x \in \mathbb{R}^{n+1}$, $\theta \in \mathbb{R}^{n+1}$ ではない ことに注意すべし

θ は x と同じ次元であるので $\theta \in \mathbb{R}^n$

Collaborative Filtering Algorithm

1. $x^{(1)}, \cdots, x^{(n_m)}, \theta^{(1)}, \cdots, \theta^{(n_u)}$ をランダムに小さい値で初期化する

2. gradient descent (or 他の方法でも可)を用いて,
$J(x^{(1)}, \cdots, x^{(n_m)}, \theta^{(1)}, \cdots, \theta^{(n_u)})$ を最小化する

for every $j=1, \cdots, n_u,\ i=1, \cdots, n_m$

$$x_k^{(i)} := x_k^{(i)} - \alpha \underbrace{\left(\sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) \theta_k^{(j)} + \lambda x_k^{(i)} \right)}_{= \frac{\partial J}{\partial x_k^{(i)}}}$$

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \underbrace{\left(\sum_{i:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) x_k^{(i)} + \lambda \theta_k^{(j)} \right)}_{= \frac{\partial J}{\partial \theta_k^{(j)}}}$$

バイアス項 $(x_0 = 1)$ が存在しないので,全ての $x_k^{(i)}$ について regularization している
θ に関しても同様

3.

Week 9　　18

[講義中のQuiz]

$x^{(1)}, \dots, x^{(n_m)}, \theta^{(1)}, \dots, \theta^{(n_u)}$ を小さいrandomな値で初期化した理由

① この手順はオプショナル(なくてもよい)。0で初期化しても動く　×

② どの問題についても gradient descent を使う時には random initialization が必要 ~~○~~ ×
放物線のようにlocal minimumが存在しない問題では初期値はall 0でもよいので ×

③ 各 i, j について $x^{(i)} \neq \theta^{(j)}$ を保証する ×

④ symetry breaking のために必要。$\underbrace{x^{(1)}, \dots, x^{(n_m)}}_{\text{それぞれ異なる}}$ ~~がそれぞれ異なる~~ を学習することを保証する ~~×~~ ○

Low Rank Matrix Factorization

$$Y = \begin{pmatrix} 5 & 5 & 0 & 0 \\ 5 & ? & ? & 0 \\ ? & 4 & 0 & ? \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 5 & 0 \end{pmatrix}$$

5x4行列

Predict ratings

$$\begin{array}{cc} & \qquad\qquad\qquad\qquad j \\ i & \begin{pmatrix} (\theta^{(1)})^T x^{(1)} & (\theta^{(2)})^T x^{(1)} & \cdots & (\theta^{(n_u)})^T x^{(1)} \\ \vdots & & \boxed{} & \vdots \\ (\theta^{(1)})^T x^{(n_m)} & & \cdots & (\theta^{(n_u)})^T x^{(n_m)} \end{pmatrix} \end{array}$$

$(\theta^{(j)})^T x^{(i)}$

$$X = \begin{pmatrix} — (x^{(1)})^T — \\ \cdots \\ — (x^{(n_m)})^T — \end{pmatrix}, \quad \Theta = \begin{pmatrix} — (\theta^{(1)})^T — \\ \cdots \\ — (\theta^{(n_u)})^T — \end{pmatrix}$$

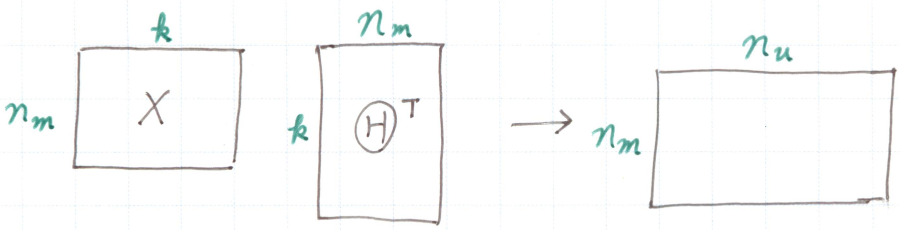
Low Rank Matrix Factorization
低ランク行列分解

$X \Theta^T$

[講義中の Quiz]

$$X = \begin{pmatrix} — (x^{(1)})^T — \\ \cdots \\ — (x^{(n_m)})^T — \end{pmatrix}, \quad \Theta = \begin{pmatrix} — (\theta^{(1)})^T — \\ \cdots \\ — (\theta^{(n_u)})^T — \end{pmatrix}$$

$$\begin{array}{cc} & n_u \\ n_m & \begin{pmatrix} (x^{(1)})^T \theta^{(1)} & \cdots & (x^{(1)})^T \theta^{(n_u)} \\ \vdots & \ddots & \vdots \\ (x^{(n_m)})^T \theta^{(1)} & \cdots & (x^{(n_m)})^T \theta^{(n_u)} \end{pmatrix} \end{array}$$ の別の表記法は?



$\therefore X \Theta^T$ が答

Finding related movies

: movie i について, feature ベクタ $x^{(i)} \in \mathbb{R}^n$ を学習した時

movie i に関連した movie j を見つけるには

small $\|x^{(i)} - x^{(j)}\|$ → j と i は似ている

Implementation detail : Mean Normalization

平均標準化

$n_u = 5$

movie \ user	user1	2	3	4	5
movie1	5	5	0	0	
2	5			0	
3		4	0		
4	0	0	5	4	
5	0	0	5		

$n_m = 5$

$n = 2$ とする. $\theta^{(5)} \in \mathbb{R}^2$

$$\min_{\substack{x^{(1)},\dots,x^{(5)} \leftarrow n_m \\ \theta^{(1)},\dots,\theta^{(5)} \leftarrow n_u}} \frac{1}{2} \underbrace{\sum_{(i,j):r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2}_{①} + \frac{\lambda}{2} \underbrace{\sum_{i=1}^{5 \,(n_m)} \sum_{k=1}^{2 \,(n)} (x_k^{(i)})^2}_{②} + \underbrace{\sum_{j=1}^{5 \,(n_u)} \sum_{k=1}^{2 \,(n)} (\theta_k^{(j)})^2}_{③}$$

user 5 は何も rating していないので ① は 0 となる. ② も 0 となる.

③ の $\frac{\lambda}{2}\left((\theta_1^{(5)})^2 + (\theta_2^{(5)})^2\right)$ を min (最小化) しようとする.

その結果 $\theta^{(5)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ となる。

$(\theta^{(5)})^T x^{(i)}$ を計算すると、どの i についても 0 となる. これは意味がない.

$$Y = \begin{pmatrix} 5 & 5 & 0 & 0 & ? \\ 5 & ? & ? & 0 & ? \\ ? & 4 & 0 & ? & ? \\ 0 & 0 & 5 & 4 & ? \\ 0 & 0 & 5 & 0 & ? \end{pmatrix}, \quad \xrightarrow[\text{(?は除く)}]{\text{平均を計算する}} \mu = \begin{pmatrix} 2.5 \\ 2.5 \\ 2 \\ 2.25 \\ 1.25 \end{pmatrix} \xrightarrow[\text{引く}]{\text{Yから}} Y = \begin{pmatrix} 2.5 & 2.5 & -2.5 & -2.5 & ? \\ 2.5 & ? & ? & -2.5 & ? \\ ? & 2 & -2 & ? & ? \\ -2.25 & -2.25 & 2.75 & 1.75 & ? \\ -1.25 & -1.25 & 3.75 & -1.25 & ? \end{pmatrix}$$

↓

$\Theta^{(j)}$, $x^{(i)}$ を学習

ユーザ j, movie i に関して

$(\theta^{(j)})^T x^{(i)} + \mu_i$

ユーザ5について $\theta^{(5)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ だったので $\underbrace{(\theta^{(5)})^T x^{(i)}}_{=0} + \mu_i = \mu_i$

Week 9　　　21

mean normalization（平均標準化）について話した.

[講義中の Quiz]

feature scaling の他のアプリケーションとは異なり, movie の rating では（max − min）で割るというスケーリングを行わなかった。

① predict されるのは実際の値なので、この種の scaling は役に立たない. ×

② 全ての rating は比較可能である（0〜5）. したがって既に同じスケールである. ○

③ 平均を引くことは範囲で割るのと数学的に同じだ　×

④ これにより全部のアルゴリズムは非常に効果的になる　×