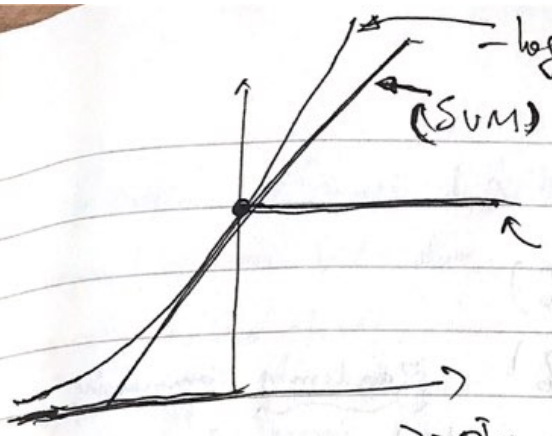


$h_\theta(x) = g(\theta^T x)$

(convex problem)

Logistic . SVM

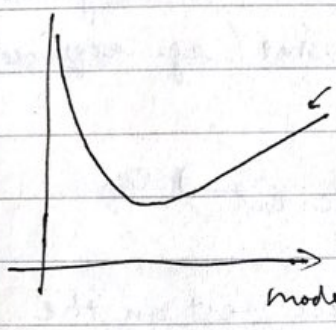
↓ approximate

min step function

(non-convex problem)

ERM Up-Hard

Model Selection



① $\theta_0 + \theta_1 x$

② $\theta_0 + \theta_1 x + \theta_2 x^2$

⋮

$\theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$

② τ - bandwidth parameter in LWLR

③ C - in SVM

$\min \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \xi_i$

$\mathcal{M} = \{M_1, M_2, M_3 \cdots\}$

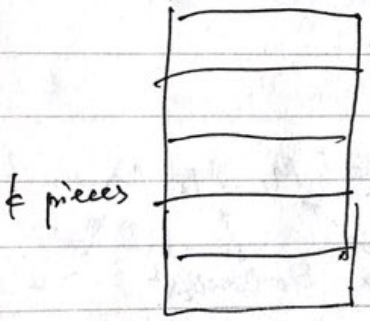
$\theta_0 + \theta_1 x$ $\theta_0 + \theta_1 x + \theta_2 x^2$ ⋯

Hold-out cross validation

- Split S into S_{train} (70%)
  $S_{c.v}$ (30%) Randomly commonly
- Train each model on S_{train}, test on $S_{c.v}$.
- Pick model with low Error on $S_{c.v}$.
- ① Retrain 100% S on best model, or
  ② Output the best model directly.

In practice, size of traing data is small in many cases. Because every training data is expensive (eg. acquired by medical experiments).

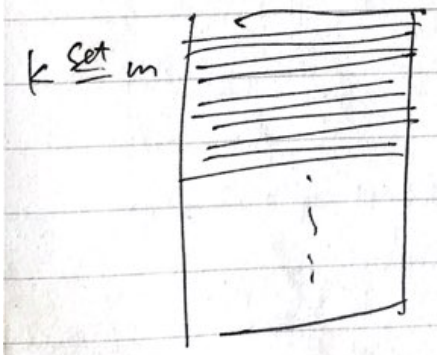
K-fold cross validation (more efficient)



- Train on k-1 pieces, test on the remaining pieces (do k times totally)
- Average over the k results
- Retraining the selected best model

* k=10 is ~~very~~ common

disadvantage: much computationally expensive



Leave-one-out C. V.

(more efficient on using data)

much more computationally expensive.

Preferred when only have little traing set

(eg. m=15)

People in Structured Risk Minimization (SRM) propose choose k depending on VC-dim, But Andrew Ng. personally tend NOT to do that.

↳ The upper bounds which proved in learning theory is very loose although in worst case of many models. (for any probability distribution). $(x, y) \in \mathcal{D}$

If plug-in constants into the formula for Learning Theory bounds, you often get extremely pessimistic estimates with # examples you need.

Eg: Logistic regression model: if have n or n+1 parameters
examples is 10 times # parameters (or tiny times)
probably in good shape (also probability).

Feature Selection

High dim. feature space. eg: 10000 features
(risk of overfitting)
(because of too many parameters in the model)

"relevant" features is much less.

With n features, there are 2^n possible subsets!

Forward search:

- Start with $\mathcal{F} = \phi$
- Repeat {
  (1) For $i = 1, \ldots, n$: try adding feature i into $\mathcal{F}$. evaluating model using cross-validation.
  (2) Set $\mathcal{F} := \mathcal{F} \cup \{$best feature found in (1)$\}$
  }
- Output best hypothesis found.

E.g.: logistic regression model: you already have 100 parameters
you may only need 10 features.
Terminate the loop when #iter exceeded 100.

"Wrapper" feature selections (computationally expensive)

But tend to work well!

Backward search:

- Start with $\mathcal{F} = \{1, \ldots, n\}$.
- Delete features one at a time.

eg: 100 training example.
10000 features. Backward search is meaningless.

Its NP-Hard to get truely best features.
All above is heuristic search.

* For text classification. Do Forward search is a heavily suffer?
too many parameters/features

"Filter" method

For each feature i: Compute some measure of how informative x_i is about y.

(eg: correlation between x_i & y)

"Corr(x_i, y)"

$$MI(x_i, y) = \sum_{x_i \in \{0,1\}} \sum_{y \in \{0,1\}} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)\, p(y)}$$

↑ Mutual Information.

$p(x_i, y)$: estimate from training data

$p(x_i)\, p(y)$: estimate from training data

[In problem set].

$$= KL\left(\left(p(x_i, y) \,\|\, p(x_i)\, p(y)\right)\right)$$

"Measure of dependence"

Pick the top k features (and choose k using C.V.)

Lecture 11

Bayesian Statistics & Regularization

Digression: Online Learning

Advice for apply ML algorithms.

Prevent overfit via regulation and keep all the parameters

Linear Regression:

maximum likelihood "Frequentist"

$$\max_{\theta} \prod_{i} p(y^{(i)} | x^{(i)}; \theta)$$

"Bayesian"

$P(\theta)$ - prior e.g. $\theta \sim \mathcal{N}(0, \tau^2 I)$

$S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$ | Calculate:

$$P(\theta|S) \propto \left(\prod_{i=1}^{m} P(y^{(i)}|x^{(i)}, \theta)\right) P(\theta)$$

($P(\theta|S)$: posterior; $P(\theta)$: Treating θ as random variable.)

To make a new prediction on x:

$$P(y|x, S) = \int_{\theta} p(y|x, \theta) P(\theta|S) d\theta.$$ ← HARD to Calculate.

$$E[y|x, S] = \int_{y} y P(y|x, S) dy$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|S)$$

$$= \arg\max_{\theta} \left(\prod_{i=1}^{m} P(y^{(i)}|x^{(i)}, \theta)\right) P(\theta).$$

To make a prediction:

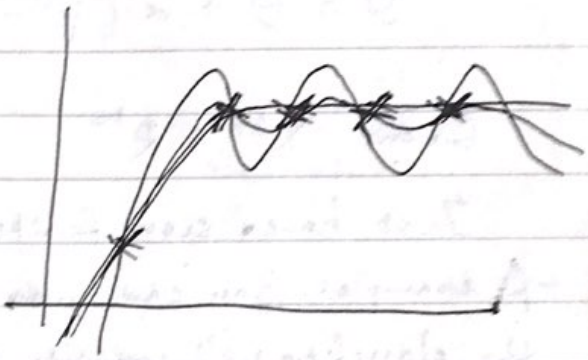
Predict $h_{\theta_{MAP}}(x) = \hat{\theta}_{MAP}^{T} x$.

$\theta \sim N(0, \tau^2 I)$ ← Set $\theta_j = 0$: same as deliminate j^{th} feature

↑ cause the effect of smaller parameter & smoother curve.

- $\Leftrightarrow$ τ smaller and smaller (closer to 0)
- $\Leftrightarrow$ curves smoother and smoother
- $\Leftrightarrow$ overfit less

[problem set 3]



e.g. LR.

$$\min_{\theta} \sum \| y^{(i)} - \theta^T x^{(i)} \|^2 + \lambda \|\theta\|^2$$

↘ make choose smaller parameter.

* With Bayesian regularization with Gaussian prior, logistic regression become very effective.

* text classification algorithm..

You can use cross-validation to choose τ^2. ?

Online Learning

(batch Learning algo. so far) ↑ different from

$x^{(1)} \to y^{(1)} \to x^{(2)} \to y^{(2)} \to x^{(3)} \to \cdots$

$\hat{y}^{(1)}$ $\hat{y}^{(2)}$ $\hat{y}^{(3)}$

web app

Total Online error:

$$\sum_{i=1}^{m} 1\{\hat{y}^{(i)} \neq y^{(i)}\}$$

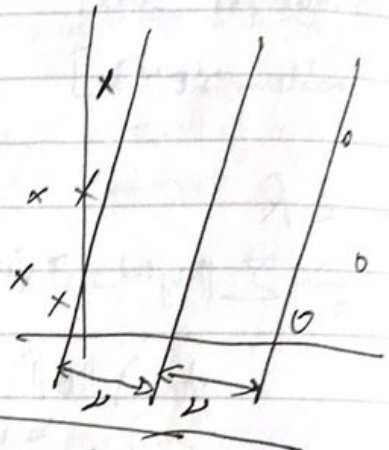
Stochastic grad. desc. algo.

E.g: Perceptron algorithm

- Initialize $\theta := 0$
- After i^{th} example, update

$$\theta := \theta + \alpha \left[y^{(i)} - h_\theta(x^{(i)}) \right] x^{(i)}$$

Even $x^{(i)} \in \mathbb{R}^{50}$



Just have seen finite number of examples, you can prove that the classifier will converge to perfect ~~seperator~~.

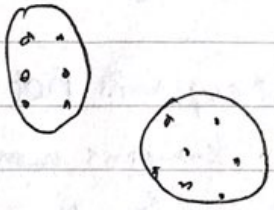
Lecture notes online gives proof (One Page)

"Advice for applying Machine Learning" [PDF]

experience ...

Lecture 12

Unsupervised Learning
- Clustering (k-means)
- Mixture of Gaussians
- Jensen's inequality
- EM (Expectation-Maximization)



"discover structure in the data"

Application in biology: cluster genes
market: cluster customers
news.google.com: cluster news
image segmentation: cluster pixels

K-means algorithm

Input: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ $\quad x^{(i)} \in \mathbb{R}^n$

1. Initialize cluster centroids
   $\mu_1, \ldots, \mu_k \in \mathbb{R}^n$
2. Repeat until convergence:
   i) Set $c^{(i)} = \arg\min_j \|x^{(i)} - \mu_j\|$.
   ii) $\mu_j := \dfrac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$

① k-means is guaranteed to converge

Distortion function

$$J(c, \mu) = \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

k-means is coordinate ascent/descent on J.

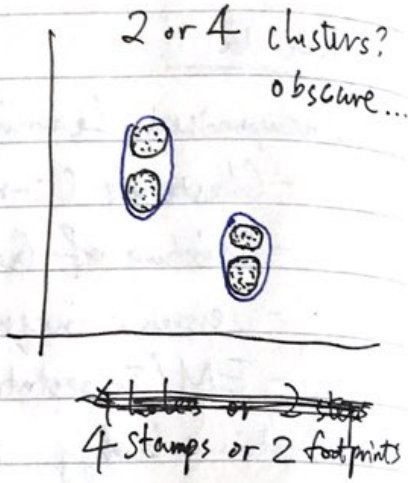
② How to choose # clusters (= k)?

Randomly choose, pick the best choose.

$J(c, \mu)$ is not convex.

③ So k-means may converge to local min. opt.

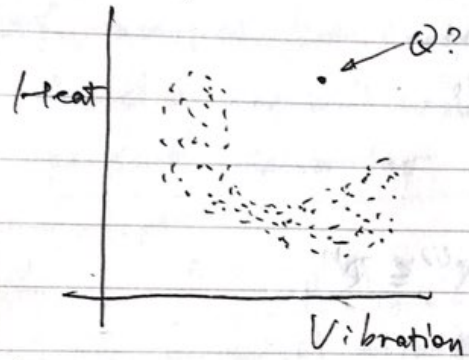
try a few times, choose the least case.



~~4 shoes or 2 shoes~~

4 stamps or 2 footprints

Density estimation



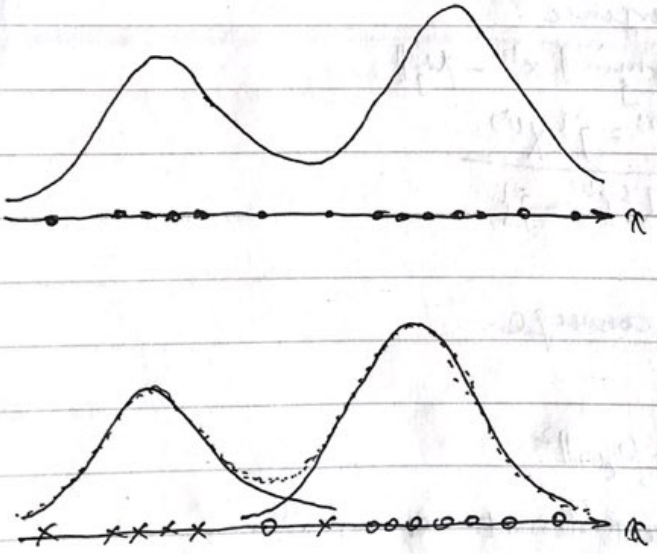
anomaly ~~abnormality~~ detection.

Q?

$P(x)$

$\{x^{(1)}, \ldots, x^{(m)}\}$.

(Aircraft Engine)



We can get the density function without make clear which two Gaussian separately.

There's a latent (hidden/unobserved) random variable z,

And $x^{(i)}, z^{(i)}$ have a joint distribution

$$p(x^{(i)}, z^{(i)}) = p(x^{(i)} | z^{(i)}) p(z^{(i)})$$

$z^{(i)} \sim \text{Multinomial}(\phi)$ (Bernoulli for 2 Gaussians)

$(\phi_j \geq 0. \quad \sum_j \phi_j = 1)$.

$x^{(i)} | z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$

Replace $y^{(i)}$ in supervised learning with latent variable $z^{(i)}$ in unsupervised learning.

If we know $z^{(i)}$'s.

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma).$$

$$\begin{cases} \phi_j = \frac{1}{m} \sum_{i=1}^{m} 1\{z^{(i)} = j\} \\ \mu_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}} \\ \Sigma_j = \cdots \quad \text{etc.} \cdots \end{cases}$$

Guess $z^{(i)}$: bootstrap procedure.

Guess → Estimate parameter (Max. likelihood) → Guess → Estimate → Guess → ···

↑ E-step ↑ M-step ↑ E-step ↑ M-step ····

EM

Repeat {

E-step: Guess values of $z^{(i)}$'s

Let $w_j^{(i)} := P(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma)$

$$= \frac{P(x^{(i)} \mid z^{(i)} = j) P(z^{(i)} = j)}{\sum_{l=1}^{k} P(x^{(i)} \mid z^{(i)} = l) P(z^{(i)} = l)}$$

$$= \frac{\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_j^{(i)})^T \Sigma_j^{-1} (x^{(i)} - \mu_j^{(i)})\right\} \cdot \phi_j}{\sum_{l=1}^{k} \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_l|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_l^{(i)})^T \Sigma_l^{-1} (x^{(i)} - \mu_l^{(i)})\right\} \cdot \phi_l}$$

M-step: Estimate parameters of Gaussian l

(GDA)

$$\begin{cases} \phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)} \\ \mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}} \\ \Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j^{(i)})(x^{(i)} - \mu_j^{(i)})^T}{\sum_{i=1}^{m} w_j^{(i)}} \end{cases}$$

← Compared to Gaussian Discriminant Analysis, Here use the probability $x^{(i)}$ comes from Gaussian j. instead of the indicator whether $x^{(i)}$ comes from Gaussian j.

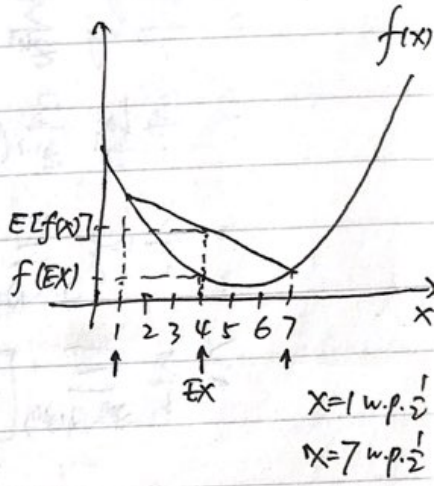
(i.e. in GDA, $w_j^{(i)} =$ "1 or 0" $= 1\{y^{(i)} = j\}$)

Jensen's inequality

Let f be a convex function (e.g. $f''(x) \geq 0$)

Let X be a random variable.

Then $f(EX) \leq E[f(X)]$.



Further, if $f''(x) > 0$ (f is strictly convex)

then $E[f(X)] = f(EX)$

$\Leftrightarrow X = EX$ w.p. 1

If $f'' \leq 0$ (f is concave)

then $f(EX) \geq E[f(X)]$. etc.

Have some model for $p(x, z; \theta)$

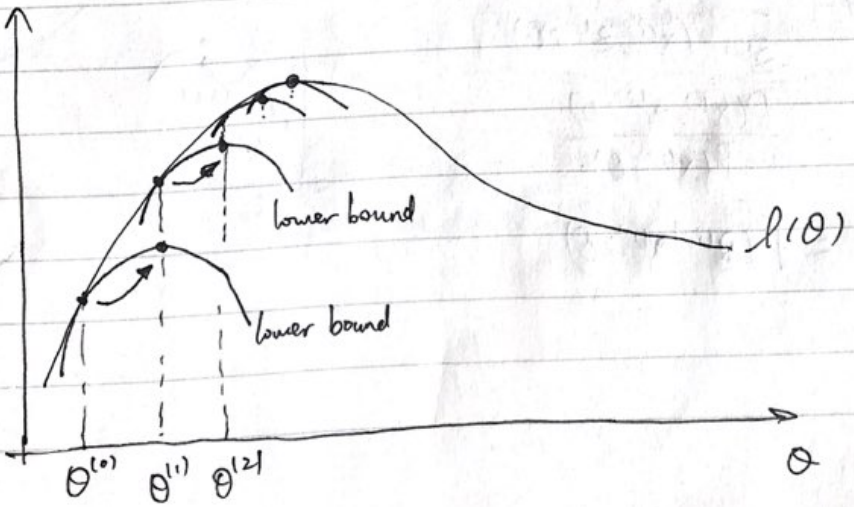
Observe only x.

Want to maximize

$$\ell(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}; \theta)$$

$$= \sum_{i=1}^{m} \log \left(\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) \right)$$

$$= \sum_{i=1}^{m} \log \left(\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \theta) \, p(z^{(i)} = l) \right)$$



$$\max_\theta \sum_i \log P(x^{(i)};\theta)$$

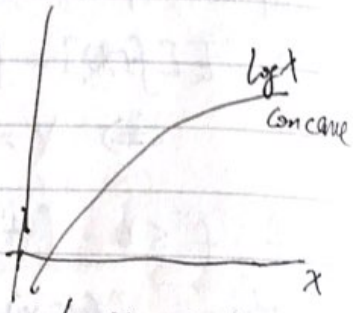
$$= \sum_i \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)};\theta)$$

$$= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \qquad \left[Q_i(z^{(i)}) \geq 0,\ \sum_{z^{(i)}} Q_i(z^{(i)}) = 1\right]$$

$$= \sum_i \log E_{z^{(i)}\sim Q_i(z^{(i)})}\left[\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$$

$$\geq \sum_i E_{z^{(i)}\sim Q_i(z^{(i)})}\left[\log \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$$

$$= \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$



$\log Ex \geq E[\log x]$

$$l(\theta) \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

You should make sure it is concave. Many case we meet it is concave at all.

Want: $\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} = \text{constant}$ (for all values of $z^{(i)}$)

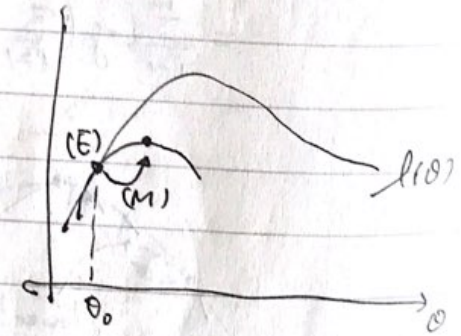
So set: $Q_i(z^{(i)}) \propto P(x^{(i)}, z^{(i)};\theta)$

$$\sum_{z^{(i)}} Q_i(z^{(i)}) = 1$$

$$Q_i(z^{(i)}) = \frac{P(x^{(i)}, z^{(i)};\theta)}{\sum_{z^{(i)}} P(x^{(i)}, z^{(i)};\theta)}$$

$$= \frac{P(x^{(i)}, z^{(i)};\theta)}{P(x^{(i)};\theta)}$$

$$= P(z^{(i)} | x^{(i)};\theta)$$



EM algorithm:

E-step:

Set $Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}; \theta)$.

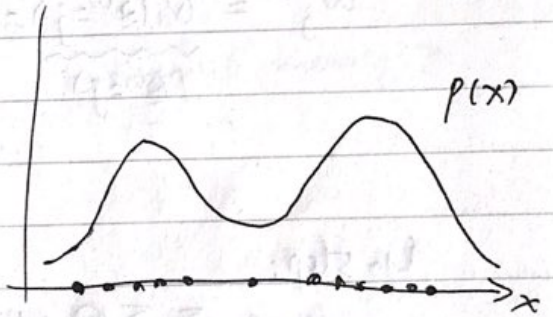
M-step:

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

Lecture 13

EM

- Mixture of Gaussian
- Mixture of Naive Bayes
- Factor analysis
- Digression: Gaussians



$\{x^{(1)}, \ldots, x^{(m)}\}$

$\max_\theta \sum_i \log P(x^{(i)}; \theta)$

$\max_\theta \sum_i \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta)$

$P(x) = \sum_z P(x|z) P(z)$

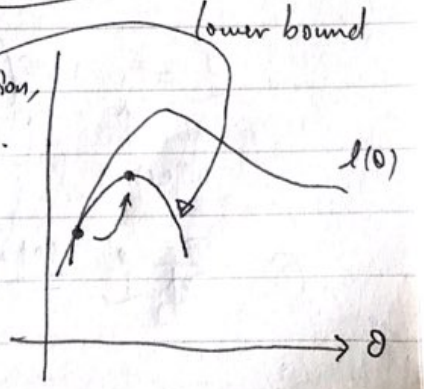
$z \sim \text{Multinomial}(\phi)$

$x|z \sim \mathcal{N}(\mu_j, \Sigma_j)$

E-step: $Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}; \theta)$

M-step: $\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$

Can easily get closed-form solution, rather than $l(\theta)$ which hard to get.



Define $J(\theta, Q) = \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$

$l(\theta) \geq J(\theta, Q)$

Co-ordinate ascent on J.

E-step: Maximize J w.r.t. Q

M-step: Maximize J w.r.t. θ

MoG (Mixture of Gaussian)

E-step:

$w_j^{(i)} = Q_i(z^{(i)}=j) = P(z^{(i)}=j \mid x^{(i)}; \phi, \mu, \Sigma)$

"$P(z^{(i)}=j)$"

$= \frac{P(x^{(i)} \mid z^{(i)}=j) P(z^{(i)}=j)}{\sum_k P(x^{(i)} \mid z^{(i)}=k) P(z^{(i)}=k)}$

$x^{(i)} \mid z^{(i)}=j \sim N(\mu_j, \Sigma_j)$

$z^{(i)} \sim$ Multinomial

M-step:

$\max_{\phi,\mu,\Sigma} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})}$

$= \sum_i \sum_j w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)\right\} \cdot \phi_j}{w_j^{(i)}}$

$w_j^{(i)} = Q_i(z^{(i)}=j)$

$P(x^{(i)} \mid z^{(i)}=j)$ $P(z^{(i)}=j)$

① $\nabla_{\mu_j}(\dots) \overset{\text{set}}{=} 0 \Rightarrow \mu_j = \frac{\sum_i w_j^{(i)} x^{(i)}}{\sum_i w_j^{(i)}}$

$\mathcal{L} = (\dots) + \beta\left(\sum_j \phi_j - 1\right)$ $\left(\sum_j \phi_j = 1 \quad \phi_j \geq 0 \text{ (ignore)}\right)$

objective; Lagrange Multipliers

② $\frac{\partial}{\partial \phi_j} \mathcal{L} \overset{\text{set}}{=} 0 \Rightarrow \phi_j = \frac{\sum_i w_j^{(i)}}{m}$

③ — — — $\Rightarrow \Sigma_j = \dots$

Text clustering: Mixture of Naive Bayes

Training set $\{x^{(1)}, \dots x^{(m)}\}$. $x^{(i)} \in \{0,1\}$ $x_j^{(i)} = 1\{\text{word } j \text{ appears in doc } i\}$

$z^{(i)} \in \{0,1\}$ (2 clusters)

$$z^{(i)} \sim \text{Bernoulli}(\phi)$$

$$p(x^{(i)}|z^{(i)}) = \prod_{j=1}^{n} p(x_j^{(i)}|z^{(i)})$$

$$p(x_j^{(i)}=1 | z^{(i)}=0) = \phi_{j|z=0}$$

E-step:

$$w^{(i)} = p(z^{(i)}=1 | x^{(i)}; \phi_{j|z}, \phi_z)$$

Initialize ϕ randomly ?!

M-step:

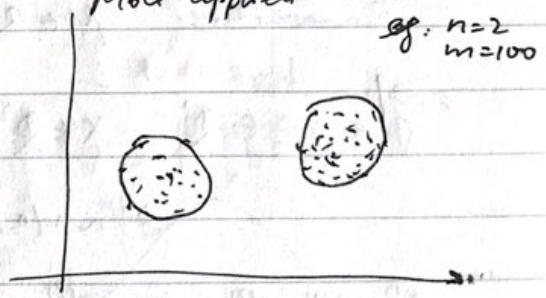
$$\phi_{j|z^{(i)}=1} = \frac{\sum_{i=1}^{m} w^{(i)} 1\{x_j^{(i)}=1\}}{\sum_{i=1}^{m} w^{(i)}}$$

$$\phi_{j|z^{(i)}=0} = \frac{\sum_{i=1}^{m} (1-w^{(i)}) 1\{x_j^{(i)}=1\}}{\sum_{i=1}^{m} (1-w^{(i)})}$$

$$\phi_{z^{(i)}} = \frac{\sum_{i=1}^{m} w^{(i)}}{m}$$

MoG applied: $m \gg n$

eg. $n=2$, $m=100$



Factor Analysis

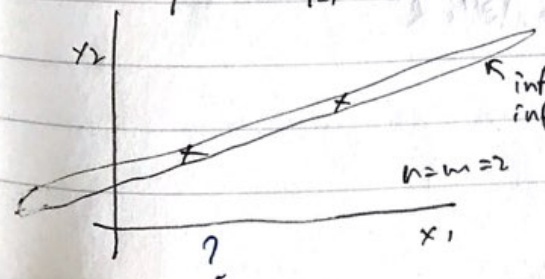
What about $n \approx m$, $n \gg m$?

Q $\{x^{(1)} \dots x^{(m)}\}$ estimate $p(x)$.

$x \sim \mathcal{N}(\mu, \Sigma)$ $\Sigma \in \mathbb{R}^{n \times n}$

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

$n \gg m \longrightarrow$ singular



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

$|\Sigma| = 0$!! Σ^{-1} undefined !

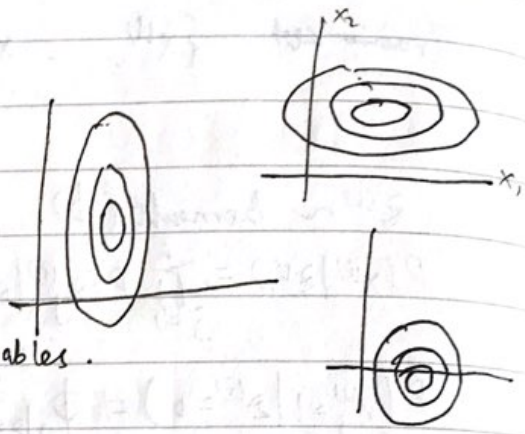
$p(x) = ?$

Not good model!

② conform Σ to be diagonal $\quad X \sim N(\mu, \Sigma)$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \sigma_n^2 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

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



Throw the correlations between variables.

Also not good!

$\Sigma = \sigma^2 I = \begin{pmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \ddots \sigma^2 \end{pmatrix}$ Even harsher constraint!

$z \sim N(0, I) \quad z \in \mathbb{R}^d$ (choose $d < n$)

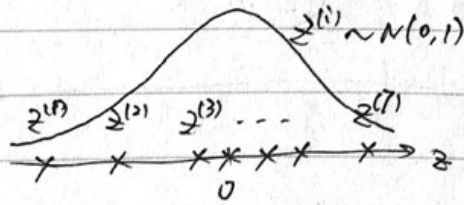
$x | z \sim N(\mu + \Lambda z, \Psi)$

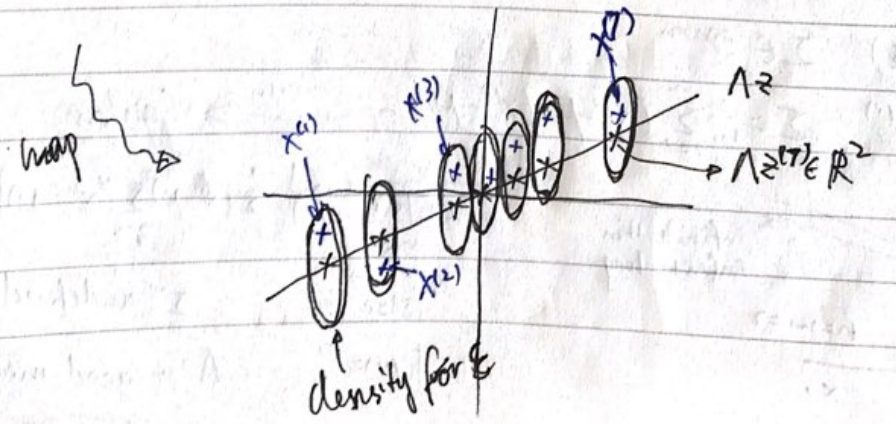
Equivalently,

$x = \mu + \Lambda z + \varepsilon$, where $\varepsilon \sim N(0, \Psi)$

Parameters: $\mu \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{n \times d}$, $\Psi \in \mathbb{R}^{n \times n}$ (diagonal)

e.g.: $z \in \mathbb{R}^1$, $x \in \mathbb{R}^2$ $\quad \Lambda = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $\Psi = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

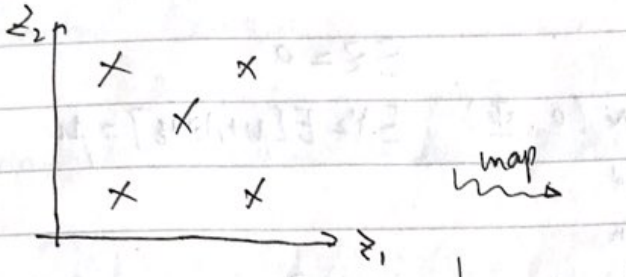




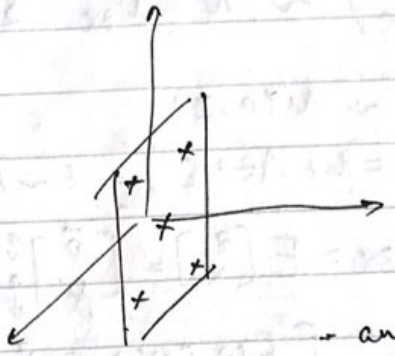
$x = \mu + \Lambda z + \varepsilon$

$\varepsilon \sim N(0, \Psi)$

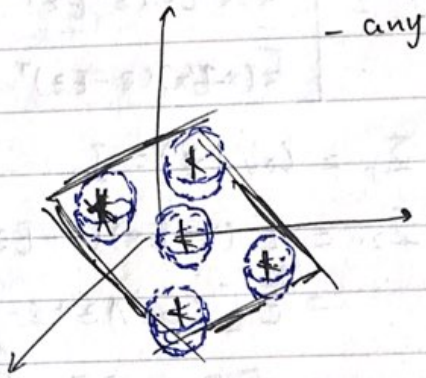
e.g. 2: $z \in \mathbb{R}^2$, $x \in \mathbb{R}^3$



map



map



- any position
- any orientation

$\mu + \Lambda z$

$x = \mu + \Lambda z + \varepsilon$

↑

3D Gaussian parallel to axis (sphere)

If $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $x_1 \in \mathbb{R}^r$ $x_2 \in \mathbb{R}^s$

$x \in \mathbb{R}^{r+s}$

$x \sim N(\mu, \Sigma)$ where $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ (r × s)

Eg: $\Sigma_{12} \in \mathbb{R}^{r \times s}$

$p(x)$

$p(x_1) = \int_{x_2} p(x_1, x_2)\, dx_2$

$x_1 \sim N(\mu_1, \Sigma_{11})$

$p(x_1 | x_2) = \dfrac{p(x_1, x_2)}{p(x_2)}$ ← $N(\mu, \Sigma)$ ← $N(\mu_2, \Sigma_{22})$

$x_1 | x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$

$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2)$

$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$

$$\begin{pmatrix} z \\ x \end{pmatrix} \sim N(\mu_{zx}, \Sigma)$$

$z \sim N(0, I)$ $\quad EZ = 0$

$x = \mu + \Lambda z + \varepsilon, \quad \varepsilon \sim N(0, \Psi) \quad EX = E[\mu + \Lambda z + \varepsilon] = \mu$

$$\mu_{zx} = E\begin{bmatrix} z \\ x \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix} \begin{matrix} \updownarrow d \\ \updownarrow n \end{matrix}$$

$$\Sigma = \left[\begin{array}{c|c} E(z-Ez)(z-Ez)^T & E(z-Ez)(x-Ex)^T \\ \hline E(x-Ex)(z-Ez)^T & E(x-Ex)(x-Ex)^T \end{array}\right]$$

$\Sigma_{11} = Cov(z) = I$

$\Sigma_{21} = E(x-Ex)(z-Ez)^T$

$= E[(\mu + \Lambda z + \varepsilon - \mu) \cdot z^T]$

$= E[\Lambda z z^T] - E[\varepsilon z^T]$

$= \Lambda E[z z^T] = \Lambda$ $\quad E[\varepsilon z^T] = 0$, $\because \varepsilon, z$ independent.

$\Sigma_{22} = E[(\mu + \Lambda z + \varepsilon - \mu)(\mu + \Lambda z + \varepsilon - \mu)^T]$

$= \cdots$

$= \Lambda\Lambda^T + \Psi$

So, $\begin{bmatrix} z \\ x \end{bmatrix} \sim N\left(\underbrace{\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}}_{\mu_{zx}}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right)$

$\{x^{(1)}, \ldots, x^{(m)}\}$ $\quad \ell(\theta) = \sum_i \log p(x^{(i)})$

$p(x^{(i)}) \quad x^{(i)} \sim N(\mu, \Lambda\Lambda^T + \Psi)$

E-step:

$Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta)$

M-step:

$$\theta := \arg\max_\theta \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} dz^{(i)}$$

Lecture 14

Factor Analysis
- EM steps

Principal Component Analysis (PCA)

$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $X \sim N(\mu, \Sigma)$ $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

Marginal: $P(x_1) = ?$ $x_1 \sim N(\mu_1, \Sigma_{11})$

Conditional: $P(x_1|x_2) = ?$ $x_1|x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$

$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$

$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

$\{x^{(1)}, \ldots, x^{(m)}\}$ $x^{(i)} \in \mathbb{R}^n$ Want $P(x)$

$z \sim N(0, I)$ $z \in \mathbb{R}^d$

$x = \mu + \Lambda z + \varepsilon$ $\varepsilon \sim N(0, \Psi)$ $x|z \sim N(\mu + \Lambda z, \Psi)$

Parameters: $\mu \in \mathbb{R}^n$ $\Lambda \in \mathbb{R}^{n \times d}$ $\Psi \in \mathbb{R}^{n \times n}$, diagonal

$\begin{pmatrix} z \\ x \end{pmatrix} \sim N(\mu_{zx}, \Sigma)$

$\begin{pmatrix} z \\ x \end{pmatrix} \sim N\left(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right)$

$P(x) = ?$

$x \sim N(\mu, \Lambda\Lambda^T + \Psi)$

$$L(\theta) = \prod_{i=1}^{m} P(x^{(i)}; \Lambda, \Psi, \mu) = \prod_{i=1}^{m} \frac{1}{(2\pi)^{\frac{n}{2}} |\Lambda\Lambda^T + \Psi|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu)\right\}$$

If maximize log-likelihood (take derivative and set to 0), cannot get analytic solution. Instead, use EM algorithm.

E-step:

Find $Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}; \theta)$

M-step:

$$\theta := \arg\max_{\theta} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} dz^{(i)}$$

① $z^{(i)}$ continous variable.

$Q_i(z^{(i)})$ continous PDF.

E-step (Computes $Q_i(z^{(i)})$):

$z^{(i)}|x^{(i)};\theta \sim N(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}})$.

where

$$\mu_{z^{(i)}|x^{(i)}} = \vec{0} - \Lambda^T(\Lambda\Lambda^T+\Psi)^{-1}(x^{(i)}-\mu)$$

"$\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu)$"

$$\Sigma_{z^{(i)}|x^{(i)}} = I - \Lambda^T(\Lambda\Lambda^T+\Psi)^{-1}\Lambda$$

"$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$"

M-step:

② $\int_{z^{(i)}} Q_i(z^{(i)}) z^{(i)} dz^{(i)}$ → unnecessary complicated:

$$\int_{z^{(i)}} \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{z^{(i)}|x^{(i)}}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(z^{(i)}-\mu_{z^{(i)}|x^{(i)}})^T \Sigma_{z^{(i)}|x^{(i)}}(z^{(i)}-\mu_{z^{(i)}|x^{(i)}})\right\} z^{(i)} dz^{(i)}$$

simpler:

$$E_{z^{(i)}\sim Q_i}[z^{(i)}] = \mu_{z^{(i)}|x^{(i)}}$$

$$E_{z^{(i)}\sim Q_i}\left[\log \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$$

$$= E_{z^{(i)}\sim Q_i}[\log P(x^{(i)}|z^{(i)};\theta)]$$ ← focus on this (only one depends on parameters)

$$+ E_{z^{(i)}\sim Q_i}\left[\log \frac{P(z^{(i)})}{Q_i(z^{(i)})}\right]$$ → many have parameter in other case; → will never have parameter (it is fixed in the E-step)

eg: $\max_{\Lambda} \sum_{i=1}^{m} E_{z^{(i)}}[\log P(x^{(i)}|z^{(i)};\Lambda,\Psi,\mu)]$

$x^{(i)}|z^{(i)} \sim N(\mu+\Lambda z, \Psi)$ $\quad \frac{1}{(2\pi)^{\frac{n}{2}}|\Psi|^{\frac{1}{2}}}\exp\{\cdots\}$

$$= \sum_{i=1}^{m} E[\text{constant} - \frac{1}{2}(x^{(i)}-\mu-\Lambda z^{(i)})^T\Psi^{-1}(x^{(i)}-\mu-\Lambda z^{(i)})]$$

$\nabla_\Lambda(\cdots) \overset{\text{set}}{=} 0$:

③ $$\Lambda = \left(\sum_{i=1}^{m}(x^{(i)}-\mu)E[z^{(i)T}]\right)\left(\sum_{i=1}^{m}E[z^{(i)}z^{(i)T}]\right)^{-1}$$

③ $\Lambda = \left(\sum_{i=1}^{m} (x^{(i)} - \mu) E_{z^{(i)}}[z^{(i)T}]\right)\left(\sum_{i=1}^{m} E[z^{(i)} z^{(i)T}]\right)^{-1}$

Expectations are w.r.t. $z^{(i)} \sim Q_i$

$\mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T$ (wrong)

$E[z^{(i)T}] = \mu_{z^{(i)}|x^{(i)}}^T$

If $z^{(i)} \sim N(\mu, \Sigma)$. $\Sigma = Ezz^T - EzEz^T$

$\Rightarrow Ezz^T = \Sigma + EzEz^T$

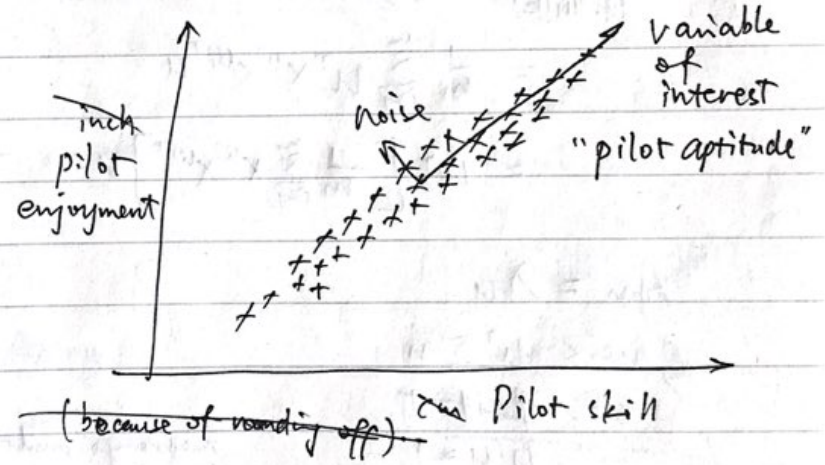
$E[z^{(i)} z^{(i)T}] = \Sigma_{z^{(i)}|x^{(i)}} + \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T$

Go to lecture notes for skipped derivation

Principal Component Analysis (PCA)

Given $\{x^{(1)}, \cdots, x^{(m)}\}$ $x^{(i)} \in \mathbb{R}^n$

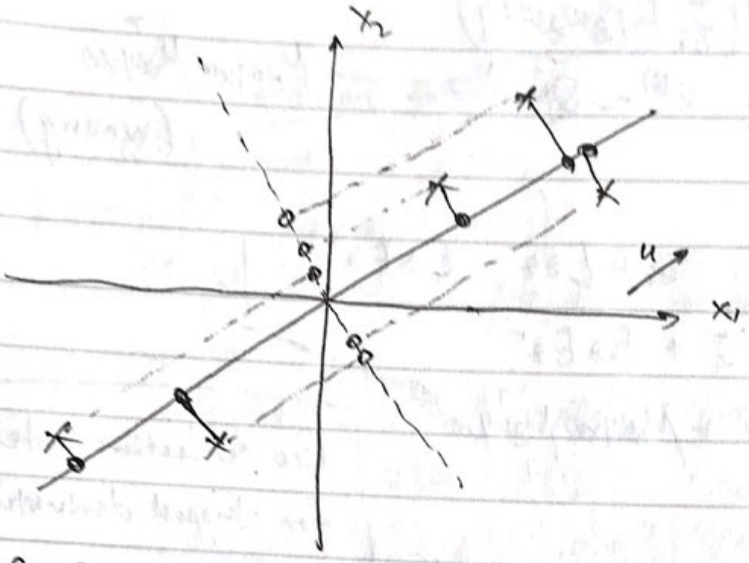
Reduce it to k-dim data ($k < n$, often $k \ll n$).



Preprocessing

1. Set $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
2. Replace $x^{(i)}$ with $x^{(i)} - \mu$ } zero out mean
3. Set $\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m} (x_j^{(i)})^2$
4. Replace $x_j^{(i)}$ with $\frac{x_j^{(i)}}{\sigma_j}$ } Normalize to unit variance

Only do this step when "different feature scales"



Objective:

Find a direction, project data onto it, get large variance.

If $\|u\|=1$, vector $x^{(i)}$ projected onto u has length $x^{(i)T}u$

Choose u:

$$\max_{u:\|u\|=1} \frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u)^2$$

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

$$= u^T\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}x^{(i)T}\right]u \quad \Rightarrow u \text{ is the principal eigenvector of } \Sigma = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}x^{(i)T}$$

$$Au = \lambda u$$

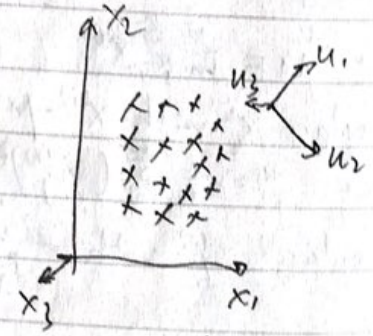
$$\begin{cases} \max\ u^T\Sigma u \\ \text{s.t. } \cancel{\|u\|=1} \\ \quad u^Tu = 1 \end{cases}$$

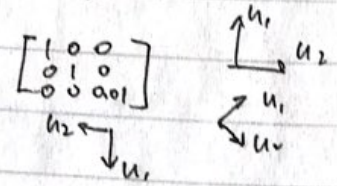
Lagrange multiplier

$$\mathcal{L}(u,\lambda) = u^T\Sigma u - \lambda(u^Tu-1)$$

$$\nabla_u \mathcal{L} = \Sigma u - \lambda u \overset{\text{set}}{=} 0;$$

If want k-dim subspace, Choose $u_1, \dots, u_k$ to be k top eigenvector of Σ.





⚠ Repeated eigenvalues: eigenvectors can rotate freely within their subspaces. Sometimes it is dangerous to look one basis at a time.