ECE M146 Introduction to Machine Learning

Prof. Lara Dolecek

ECE Department, UCLA

Today’s Lecture

- Overview of ML terminology

- Math review

Today's Lecture

- Overview of ML terminology

- Math review

Why Take Intro to Machine Learning Course ?

- Goals of this course:

1. Learn about most popular ML algorithms and frameworks
2. Develop mathematical understanding as well as intuition for these techniques
3. Test and implement these methods on various examples

What is Machine Learning ?

- Build a system (model/algorithm) based on training data to make inferences on testing data.

- "Machine" part is what specifies this system (model/algorithm). Optimize parameters of the system with respect to the training data.

Real Life Examples and Applications Abound!

- Medicine – from symptoms to diagnosis
- Transportation – automation of cars
- Privacy and authentication -- facial recognition
- Stock market
- Consumer behavior – Amazon
- …and many more!

Human vs. Machine Learning

- Human learning – uses complex hypotheses set, context, generative modeling

- Machine learning – pattern recognition equipped with a mathematical model for it. Can need a lot of training data for tasks that are “simple” for humans.

Classification of ML Algorithms

- Broad classification of ML algorithms, based on what is available at the training time:
- Training

$\tilde{x} \xrightarrow{f} \tilde{y}$ f is unknown

- Build a Model

g proxy for f.

- Testing

$x \xrightarrow{g} y$

- <u>Supervised Learning</u>: at training time, we have access both to inputs and their labeled outputs.

Classification of ML Algorithms

- Broad classification of ML algorithms, based on what is available at the training time:
- Training
- Build a Model
- Testing

- Unsupervised Learning: at training time, we have access only to input data, but not their labeled outputs.

Supervised Learning

1. Classification

2. Regression

Supervised Learning

1. Classification
    - Example: automated digit sorter for zip codes in hand written mail

0 1 2 3 4 5 6 7 8 9

5 7

2. Regression

Supervised Learning

1. Classification
    - Example: automated digit sorter for zip codes in hand written mail

2. Regression
    - Example: given the recent housing sales in zip code 90210, predict the sale value of a 4-bedroom house there.

Classification vs. Regression

- Key difference: at test time,
- in classification, determine the value among a finite set of choices that appeared in the training set.

Classification vs. Regression

- Key difference: at test time,
- in classification, determine the value among a finite set of choices that appeared in the training set.

- In regression, predict/assign a (possibly new) real value to the test point, based on the input-output relationship in the training data.

Algorithms for Supervised Learning

- Perceptron
- Logistic Regression
- Decision Trees
- Linear-least squares
- K-Nearest Neighbors
- Support Vector Machines
- Naïve Bayes Classifier
- Gaussian Discriminant Analysis

Algorithms for Supervised Learning

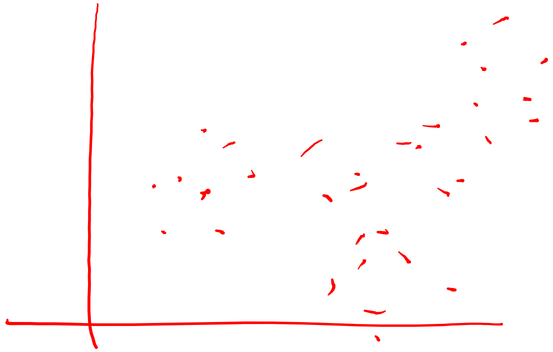
- Perceptron
- Logistic Regression
- Decision Trees
- Linear-least squares
- K-Nearest Neighbors
- Support Vector Machines
- Naïve Bayes Classifier
- Gaussian Discriminant Analysis

- Linear vs. non-linear decision boundary
- Probabilistic vs. non-probabilistic setting
- Discriminative vs. generative
- On-line vs. batch updates

Unsupervised Learning

- At training time, we do not have access to labels, only the data.
- Main settings:

1. Clustering



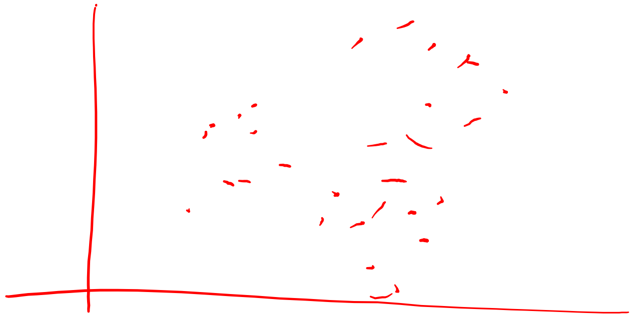
2. Projections/dimensionality reduction

3. Density Estimation

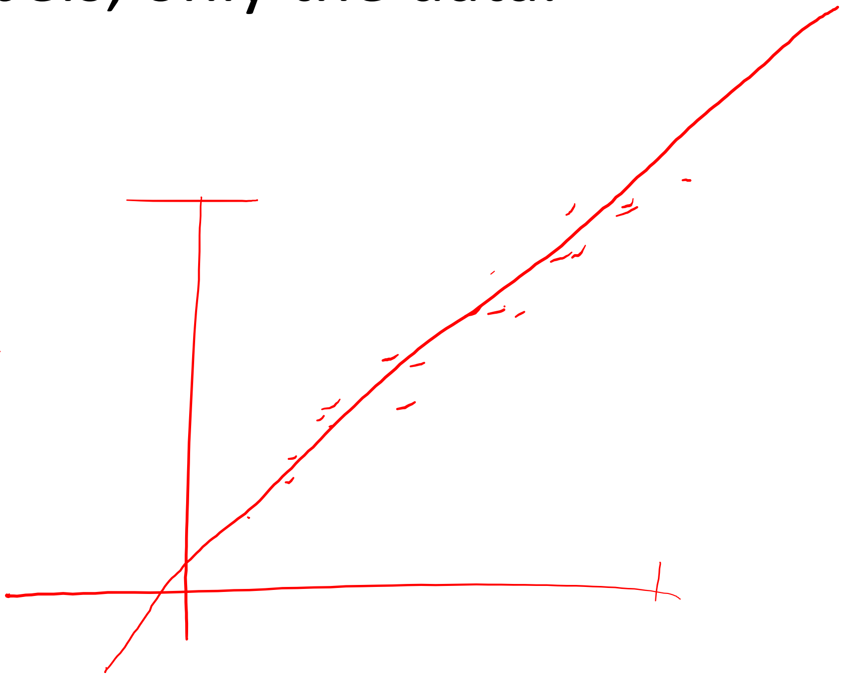
Unsupervised Learning

- At training time, we do not have access to labels, only the data.
- Main settings:

1. Clustering
   - K-means clustering



2. Projections/dimensionality reduction
   - Principal Component Analysis (PCA)



3. Density Estimation
   - MLE estimation

Reinforcement Learning

- Instead of the the x → y relationship, we observe (x,z), where z is partial information about y.

- Instead of providing "good" training examples, as in supervised learning, algorithm needs to discover suitable actions by trial and error for maximum reward.

Reinforcement Learning

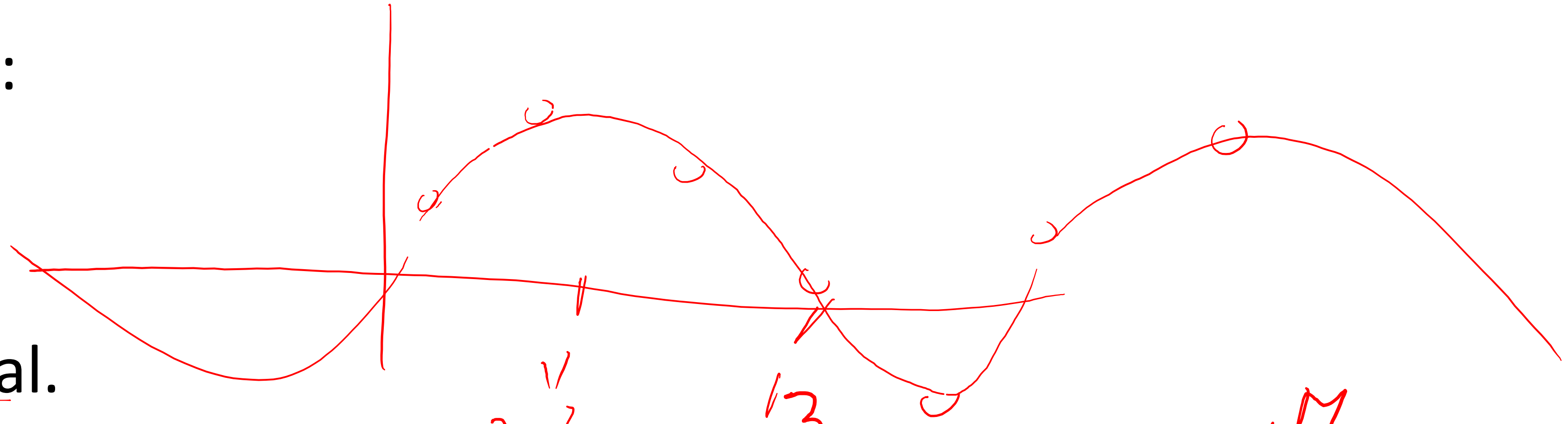
- Instead of the the x → y relationship, we observe (x,z), where z is partial information about y.

- Instead of providing “good” training examples, as in supervised learning, algorithm needs to discover suitable actions by trial and error for maximum reward.

- There is a tradeoff between exploration (attempt to discover new actions) and exploitation (maximize rewards based on available actions).

Polynomial curve fitting

- Let’s use it to solve a regression problem; will highlight important issues as well.
- Suppose we want to fit a polynomial to a sin(∏x) function (allow for random noise).
- Polynomial function:



- Degree M polynomial.

$$f(x, \underline{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_M x^M$$

Polynomial curve fitting

- How to design this function ?
- It depends on what is available.

Polynomial curve fitting

- How to design this function ?
- It depends on what is available.

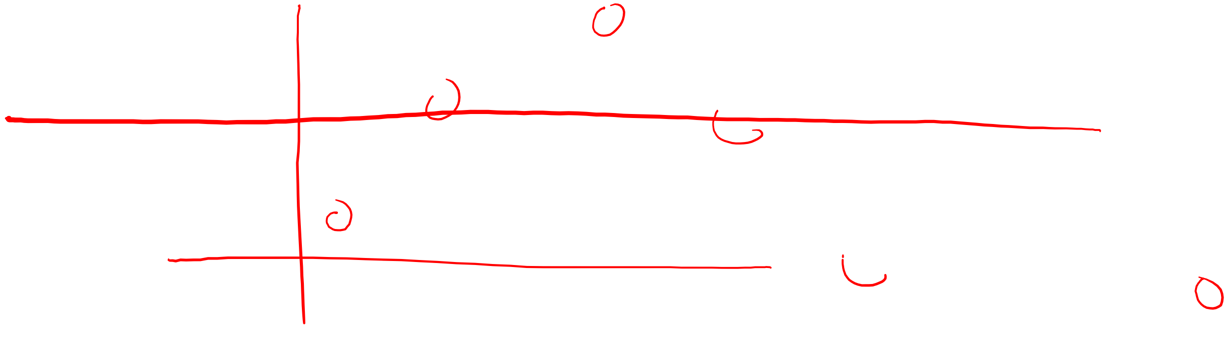
- **Idea #1: Given the training set, find the best fit.**

- Ok, let's see how. Specify the degree of the polynomial, M.

Polynomial curve fitting – idea #1

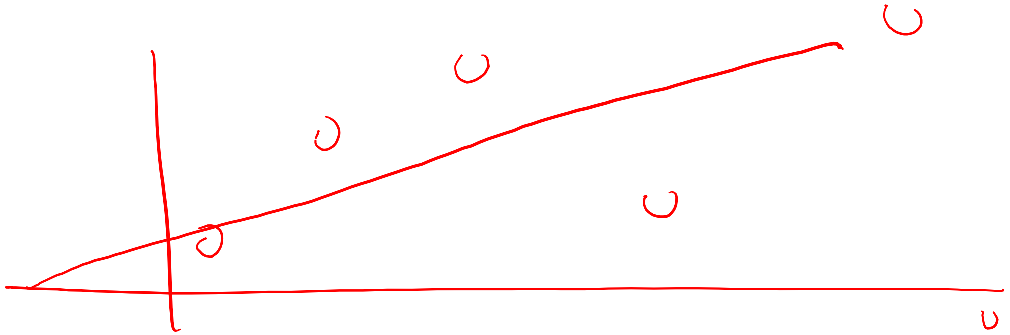
- Degree M = 0
  - Fit a horizonal line

$f = w_0$



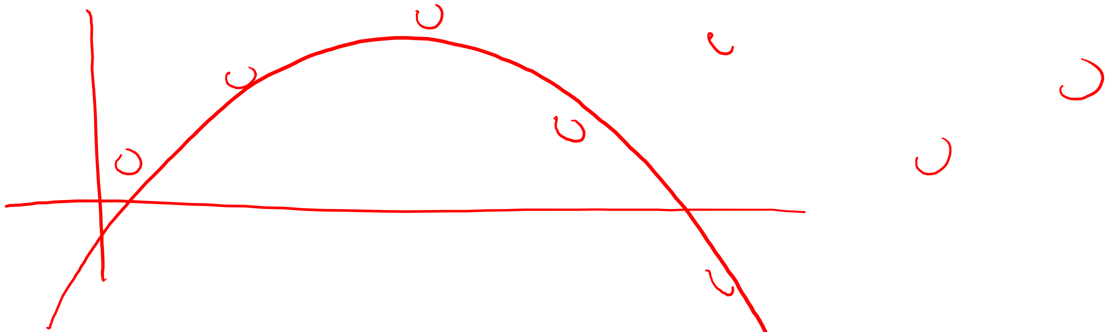
- Degree M = 1
  - Fit a line

$f = w_0 + w_1 x$



- Degree M = 2
  - Fit a quadratic

$f = w_0 + w_1 x + w_2 x^2$



Polynomial curve fitting – idea #1

- Fitting a best curve means to find a polynomial of the given degree such that the error on the training set is minimized

$$\text{Error} = \sum_{i=1}^{N} (\tilde{y}_i - f(x_i, \underline{w}))^2$$

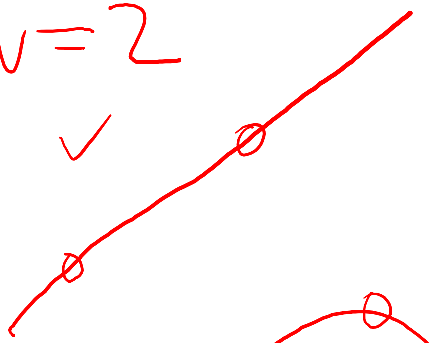
- Here, N denotes the number of training examples.

Polynomial curve fitting – idea #1

- Fitting a best curve means to find a polynomial of the given degree such that the error on the training set is minimized

fit Line

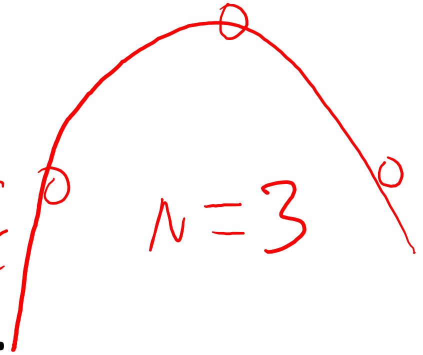
N=2 ✓



- Here, N denotes the number of training examples.

fit quadratic

N=3



- Can we get the error to be zero ?
- Yes! Can always fit a polynomial of degree N-1 to N points.

Not so fast.

- Observe what we really want is a model that **generalizes well to unseen (testing) data**.
- Why is M = N-1 a bad idea ?

Not so fast.

- Observe what we really want is a model that **generalizes well to unseen (testing) data**.
- Why is M = N-1 a bad idea ?
- Totally unpredictable behavior outside the training set!
- This is known as **overfitting**.

Solution: Regularization

- This is idea #2
- Error formula: $\text{Error} = \sum_{i=1}^{N} (\tilde{y}_i - f(\tilde{x}_i, w))^2 + \underbrace{\lambda \cdot \|w\|^2}$

- Term with the parameter lambda Λ penalizes large magnitude values; value of lambda specifies by how much (non-negative parameter).

Today's Lecture

- Overview of ML terminology

- Math review

- Probability
- Linear Algebra
- Optimization (later)

Probability

- Random Variables

$X : \Omega \to \mathbb{R}$

Ω is set of outcomes

Probability

- Random Variables
- Conditional Probability and Bayes Rule

Probability

- Random Variables
- Conditional Probability and Bayes Rule
- Examples of Important Random Variables: Bernoulli, Uniform, Gaussian

Probability

- Random Variables
- Conditional Probability and Bayes Rule
- Examples of Important Random Variables: Bernoulli, Uniform, Gaussian
- Maximum Likelihood Estimation (MLE)

Example: Drawing balls from boxes

- Set up: box #1 box #2

box #1: r_1 RED, b_1 BLUE

box #2: r_2 RED, b_2 BLUE

- Pick a box at random. Draw a ball from this box.
- Suppose the drawn ball is red.

- What is the probability that the drawn ball is drawn from box #2?

Example: Drawing balls from boxes

- Let X be the random variable denting the index of the box.
- Values of X ?

$X \in \{1, 2\}$

$P(\text{ball is red}) = P(\text{ball is red} \mid X=1) \cdot P(X=1) + P(\text{ball is red} \mid X=2) \cdot P(X=2)$

- Conditional probability:

$$P(X=2 \mid \text{ball is red}) = \frac{P(X=2, \text{ball is red})}{P(\text{ball is red})} = \frac{P(\text{ball is red} \mid X=2) \cdot P(X=2)}{P(\text{ball is red})}$$

$$= \frac{\frac{r_2}{r_2+b_2} \cdot \frac{1}{2}}{\frac{r_1}{r_1+b_1} \cdot \frac{1}{2} + \frac{r_2}{r_2+b_2} \cdot \frac{1}{2}}$$

Review: Bayes Rule

- $P(A|B) = \dfrac{P(B|A)\,P(A)}{\underline{P(B)}}$

$$\underline{P(B)} = \sum_{k=1}^{K} P(B|A_k)\cdot P(A_k)$$

total probability law

$A_1, A_2, \ldots A_K$ form a partition

$$\bigcup_{k=1}^{K} A_k = S \qquad ② \; A_j \cap A_k = \emptyset \quad j \neq k$$

Review: Bernoulli Random Variable

- X is a Bernoulli RV if it takes on value "1" with probability p, and value "0" with probability (1-p).
- Boxes and balls example, revisited.
- Pick box #1 with probability q, and pick box #2 with probability (1-q). Draw a ball from this box.
- Suppose the drawn ball is red.
- What is the probability that the drawn ball is drawn from box #2?

Example, continued

- Conditional probability:

$$P(X=2 \mid \text{saw is red}) = \frac{\frac{r_2}{r_2+b_2}\cdot(1-q)}{\frac{r_1}{r_1+b_1}\cdot q + \frac{r_2}{r_2+b_2}\cdot(1-q)}$$

Review: Uniform Random Variable

- Discrete

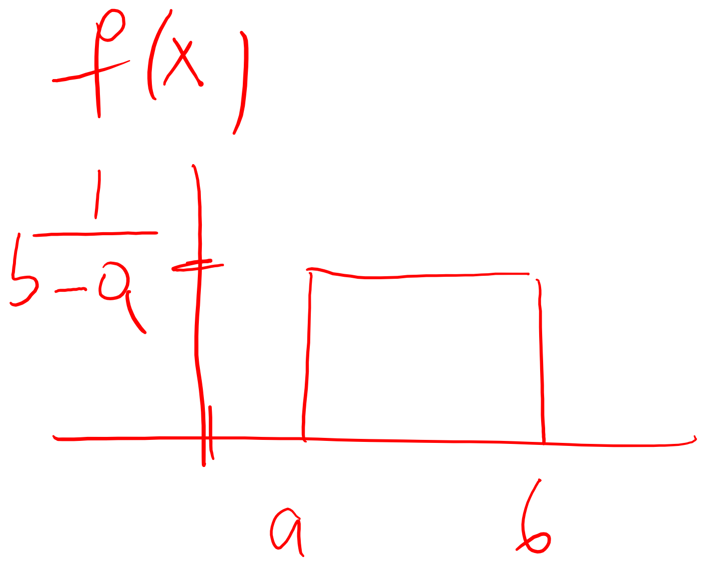
$X \in \{1, 2, \ldots, L\}$

$P(X=i) = \frac{1}{L}$

$1 \le i \le L$

- Continuous

$X \in [a, b]$



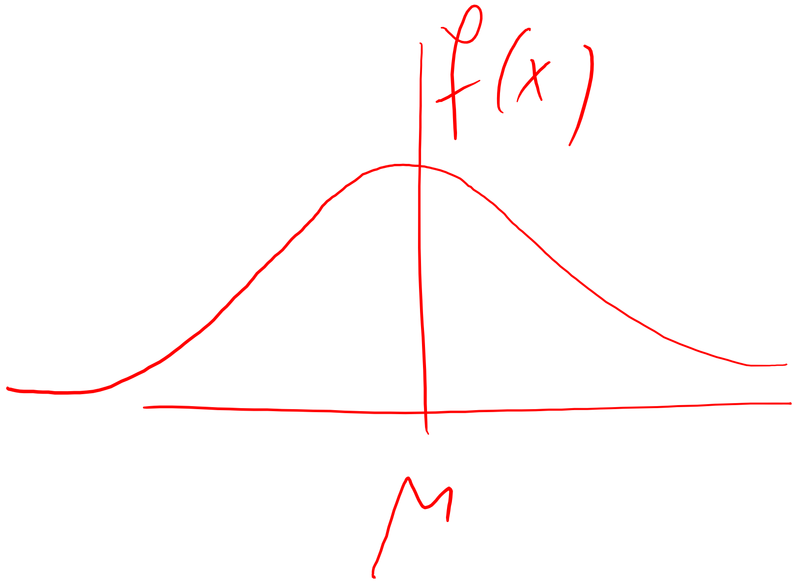
$\int f(x) = 1$

Review: Gaussian Random Variable

- A random variable X is said to be Gaussian if its pdf has the following format: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$-\infty < x < +\infty$$



- Interpretation for the mean and variance

$$\mu = E[X] \qquad VAR(X) = \sigma^2$$

Review: mean and variance of a RV

- Discrete RV

$$E[X] = \sum_{k} x_k \cdot P(X = x_k)$$

$$E[g(X)] = \sum_{k} g(x_k) \cdot P(X = x_k)$$

$$g(x) = x^2$$

- Continuous RV

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f_X(x)\, dx$$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)\, f_X(x)\, dx$$

$$VAR(X) = E[X^2] - (E[X])^2$$

Multivariate Gaussian RV

$$\mathrm{cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$$

- A vector RV X is said to be multivariate jointly Gaussian if its pdf has the following format:

$$f_X(x) = \frac{1}{(2\pi)^{D/2}\sqrt{\det(\Sigma)}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right\}$$

- Here, D is the dimension of X.
- Interpretation for the mean and covariance matrix. Σ

$$\mu = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_D] \end{bmatrix}$$

μ

Σ is $D \times D$ matrix

$$\Sigma_{ij} = \mathrm{cov}(x_i, x_j)$$

More on variance and covariance

$$\Sigma = \begin{bmatrix} VAR(x_1) & COV(x_1, x_2) & \cdots & COV(x_1, x_D) \\ COV(x_2, x_1) & VAR(x_2) & & \\ \vdots & & \ddots & \\ COV(x_D, x_1) & & \cdots & VAR(x_D) \end{bmatrix}$$

$\forall z \neq 0$

$z^T \cdot \Sigma \cdot z > 0$

P.D.

$cov(x_i, x_j) = cov(x_j, x_i)$

Σ is symmetric

For jointly gaussian Σ^{-1} exists

in general, Σ is positive semi-definite matrix

For jointly gaussian, Σ is positive definite matrix

Linear Algebra and Gaussian RV

- Covariance matrix ∑.
- The inverse of ∑ is Σ^{-1}

$\Sigma \cdot \Sigma^{-1} = I$

- Determinant of ∑ is det(∑).

$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$

- For jointly Gaussian, matrix ∑ is positive definite. Inverse exists and determinant is positive.

Special case: Covariance Matrix is Diagonal

+ same variance

$$f_X(x) = \frac{1}{(2\pi)^{D/2} \cdot (\sigma^2)^{D/2}} \cdot \exp\left\{ -\frac{1}{2} \frac{1}{\sigma^2} \sum_{i=1}^{D} (x_i - \mu_i)^2 \right\}$$

$$\exp(\Sigma) \to \Pi \exp$$

$$\Sigma = \begin{bmatrix} \sigma^2 & & 0 \\ 0 & \sigma^2 & \\ 0 & & \ddots & \\ & & & \sigma^2 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & & 0 \\ & \frac{1}{\sigma^2} & \\ 0 & & \ddots & \\ & & & \frac{1}{\sigma^2} \end{bmatrix}$$

$$\det(\Sigma) = (\sigma^2)^D$$

$$\underbrace{(x-\mu)^T}_{v} \cdot \underbrace{\Sigma^{-1}}_{M} \underbrace{(x-\mu)}_{v}$$

Review: Linear algebra

- Recall rules for taking the transposes

$$x \cdot A = y \quad /T$$

$$\underline{1 \times D} \quad \underline{D \times n} \qquad 1 \times n$$

$$(X \cdot A)^T = y^T \qquad A^T \cdot X^T = y^T$$

$$n \times \cancel{D} \quad \cancel{D} \times 1 \qquad n \times 1$$

Review: Linear algebra

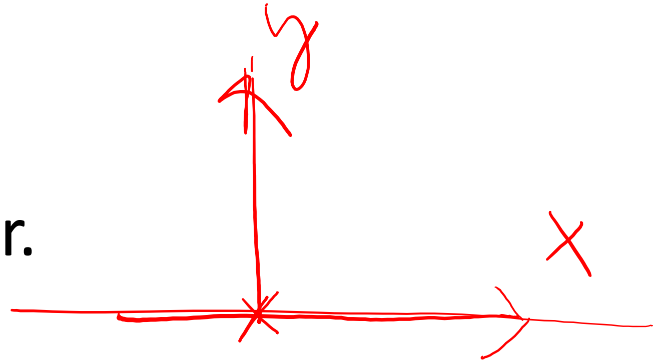
- Vector projections.
- Consider vector x of dimension D.
- We write || x || for vector norm
- L1 norm $\sum_{i=1}^{D} |x_i|$
- L2 norm $\sqrt{\sum_{i=1}^{D} x_i^2}$

Review: linear algebra

- Inner product of vectors x and y.

$$x^T \cdot y = \sum_{i=1}^{D} x_i y_i$$

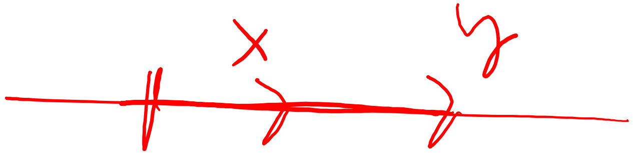
- If $x^T y = 0$, x and y are perpendicular.



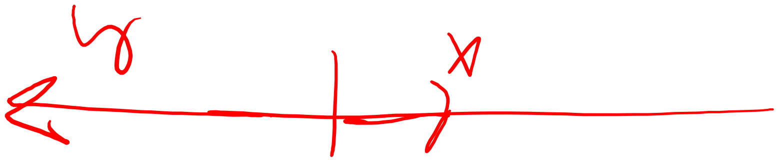
- When is the projection maximized? When y = a*x.

$$x^T y = \|x\| \cdot \|y\| \cdot \cos\theta$$

$a > 0$



$a < 0$



Parameter Estimation

- Suppose we sample from an iid Gaussian with known variance and unknown mean.

$$f_X(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Goal is to provide the best estimate of the mean based on the sampled data. N

$$f(x) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Maximize this likelihood

Parameter estimation – ctd.

take the log so that product becomes a sum.

$$N \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu} = 0$$

$$\frac{\partial}{\partial\mu}\left(N\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right) = 0$$

$$\frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right) = -\frac{1}{\cancel{2}\sigma^2}\sum_{i=1}^{N}\cancel{2}(x_i-\mu)(-1) = 0$$

$$\sum_{i=1}^{N}(x_i-\mu) = 0$$

$$\sum_{i=1}^{N}x_i = N\cdot\mu$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}x_i$$