CSE 575: Statistical Machine Learning

Jingrui He

CIDSE, ASU

MLE
Linear Regression

Your first consulting job

- A billionaire from Tempe asks you a question:
  - He says: I have a thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times:
  - You say: The probability is:
  - **He says: Why???**
  - You say: Because…

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data:** Observed set *D* of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?
- ***MLE***: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Your First Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ \ln P(\mathcal{D} \mid \theta) = 0}$

How Many Flips Do I Need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Humm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

Simple Bound (based on Hoeffding's inequality)

- For $N = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

- Let θ^* be the true parameter, for any $\varepsilon > 0$:

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

PAC Learning

- PAC: Probably Approximately Correct
- Billionaire says: I want to know the thumbtack parameter θ, within $\varepsilon = 0.1$, with probability at least $1-\delta = 0.95$. How many flips?

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\; \leq \;\; 2e^{-2N\epsilon^2}$$

What about prior

- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

- Rather than estimating a single θ, we obtain a distribution over possible values of θ

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Bayesian Learning for Thumbtack

$$P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta)P(\theta)$$

- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

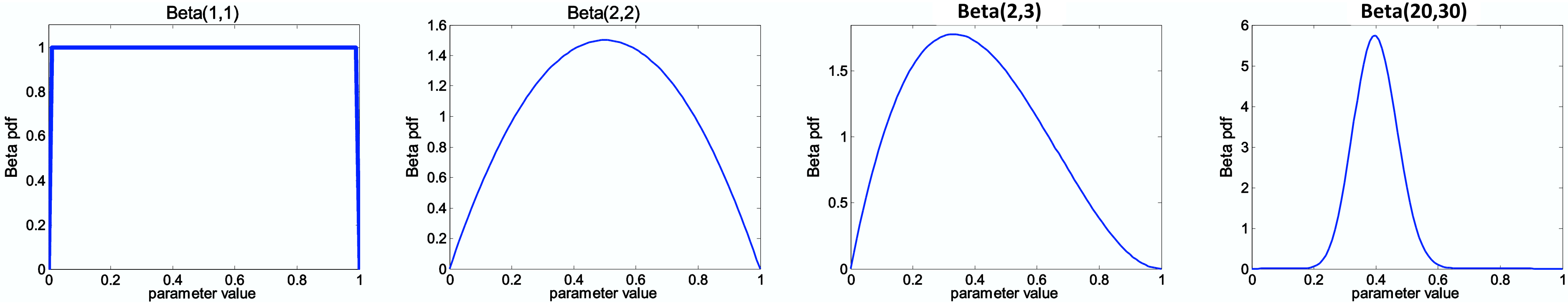
- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Prior/posterior: same probability distribution family
  - **For Binomial, conjugate prior is Beta distribution**

Beta Prior Distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Mean: $\dfrac{b_H}{b_H + b_T}$

Mode: $\dfrac{b_H - 1}{b_H + b_T - 2}$



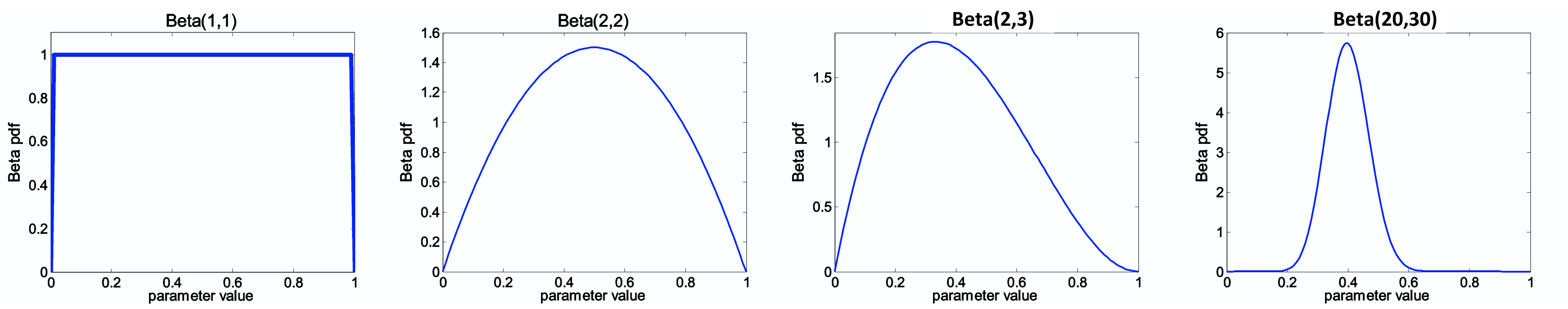
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta)P(\theta)$

Posterior Distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails

- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Using Bayesian posterior

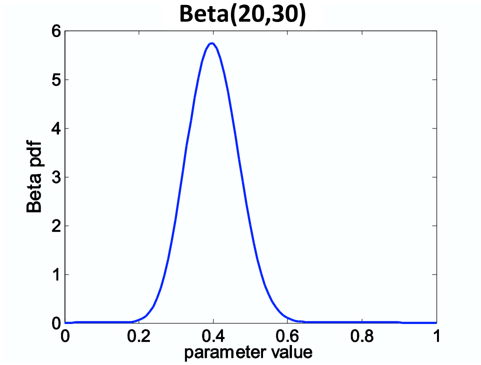
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter:

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

  - Integral is often hard to compute



MAP: Maximum a Posteriori Approximation

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

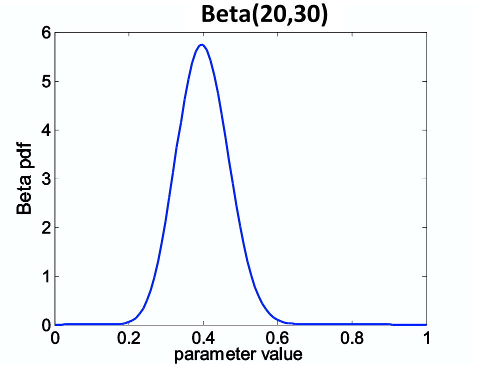
- As more data is observed, Beta is more certain

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_\theta P(\theta \mid \mathcal{D}) \quad E[f(\theta)] \approx f(\widehat{\theta})$$

MAP for Beta Distribution

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$



- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) =$$

- Beta prior equivalent to extra thumbtack flips
- As $N \to \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

What About Continuous Variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Some Properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$ Independence?

Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean
  - Variance

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu, \sigma) &= \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

Your Second Learning Algorithm: MLE for Mean of a Gaussian

- What's MLE for mean?

$$\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) \;=\; \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

Properties of MLE for Mean

- Under certain conditions, MLE is consistent

$$\hat{m}_{MLE} \xrightarrow{P} m^*$$

- Asymptotic Normality: let $se = \sqrt{Var_m(\hat{m}_{MLE})}$.
  Under regularity conditions,

$$\frac{\widehat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1) \qquad se \approx \sqrt{1/I_n(\theta)}$$

Fisher Information

MLE for Variance

- Again, set derivative to zero:

$$\begin{aligned}
\frac{d}{d\sigma}\ln P(\mathcal{D}\mid\mu,\sigma) &= \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]\\
&= \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}\right]-\sum_{i=1}^{N}\frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]
\end{aligned}$$

Learning Gaussian Parameters

- MLE: $$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

$$\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

Bayesian Learning of Gaussian Parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$

MAP for Mean of Gaussian

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}} \qquad P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\boxed{\frac{d}{d\mu}\left[\ln P(\mathcal{D} \mid \mu) P(\mu)\right] \;=\; \frac{d}{d\mu}\left[\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)\right]}$$

Frequentist Statistics

- Data are random
- Estimators are random because they are functions of data
- Parameters are fixed, unknown constants not subject to probabilistic statements
- Procedures are subject to probabilistic statements, for example 95% confidence intervals trap the true parameter value 95% of the time
- Classifiers, even learned with deterministic procedures, are random because the training set is random
- PAC bound is frequentist

Bayesian Statistics

- Probability refers to degree of belief
- Inference about a parameter θ is by producing a probability distributions on it
- Starts with prior distribution p(θ)
- Likelihood function p(x | θ), a function of θ not x
- After observing data x, one applies the Bayes rule to obtain the posterior
- Prediction by integrating parameters out:

$$p(x \mid Data) = \int p(x \mid \theta) p(\theta \mid Data) d\theta$$

Prediction of Continuous Variables

- Billionaire says: Wait, that’s not what I meant!
- You says: Chill out, dude.
- He says: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- You say: **I can regress that…**

The Regression Problem

- **Instances:** <$\mathbf{x}_j$, t_j>
- **Learn:** Mapping from x to t(**x**)
- **Hypothesis space:**
  - Given, basis functions
  - Find coeffs **w**={w_1,…,w_k}

$$H = \{h_1, \ldots, h_K\}$$

$$\underbrace{t(\mathbf{x})}_{\text{data}} \approx \widehat{f}(\mathbf{x}) = \sum_i w_i h_i(\mathbf{x})$$

  - Why is this called linear regression???
    - model is linear in the parameters

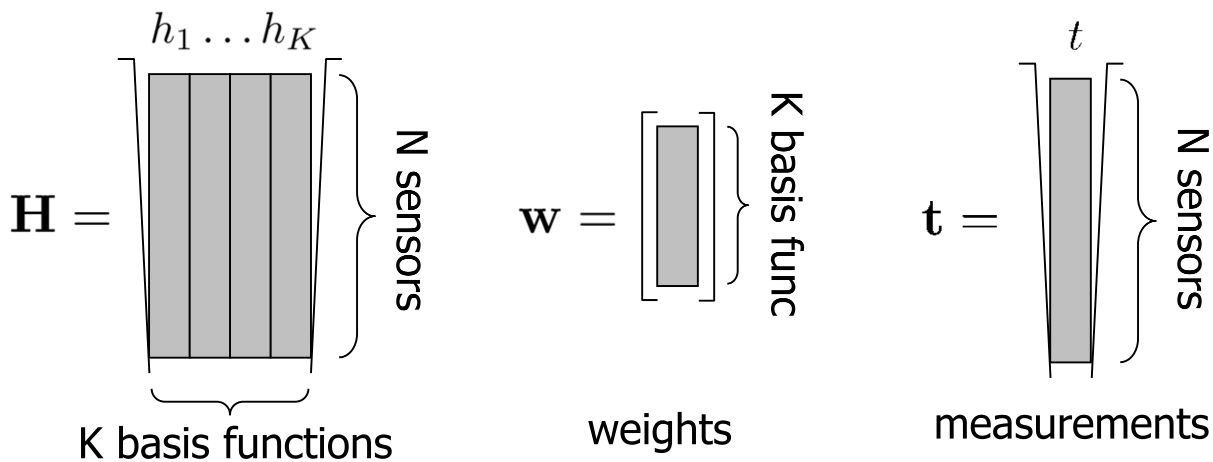
- **Precisely, minimize the residual squared error:**

$$\mathbf{w}^* \quad = \quad \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

Regression in Matrix Notation

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{(\mathbf{Hw} - \mathbf{t})^T(\mathbf{Hw} - \mathbf{t})}_{\text{residual error}}$$



Regression Solution: Matrix Operations

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{(\mathbf{H}\mathbf{w} - \mathbf{t})^T (\mathbf{H}\mathbf{w} - \mathbf{t})}_{\text{residual error}}$$

$$\text{solution: } \mathbf{w}^* = \underbrace{(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}}_{\mathbf{A}^{-1}} \underbrace{\mathbf{H}^{\mathrm{T}}\mathbf{t}}_{\mathbf{b}} = \mathbf{A}^{-1}\mathbf{b}$$

$$\text{where } \mathbf{A} = \mathbf{H}^{\mathrm{T}}\mathbf{H} = \underbrace{\begin{bmatrix} \square & \square & \square & \square \\ \square & \square & \square & \square \\ \square & \square & \square & \square \\ \square & \square & \square & \square \end{bmatrix}}_{\substack{\text{k×k matrix} \\ \text{for k basis functions}}} \qquad \mathbf{b} = \mathbf{H}^{\mathrm{T}}\mathbf{t} = \underbrace{\begin{bmatrix} \square \\ \square \\ \square \\ \square \end{bmatrix}}_{\text{k×1 vector}}$$

But, Why?

- Billionaire (again) says: Why sum squared error???
- You say: Gaussians, Dr. Gateson, Gaussians…

- Model: prediction is linear function plus Gaussian noise
  - $t = \sum_i w_i h_i(\mathbf{x}) + \varepsilon$

- Learn **w** using MLE

$$P(t \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\left[t - \sum_i w_i h_i(\mathbf{x})\right]^2}{2\sigma^2}}$$

Maximizing Log-likelihood

Maximize:

$$\ln P(\mathcal{D} \mid \mathbf{w}, \sigma) = \ln \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{j=1}^{N} e^{\frac{-[t_j - \sum_i w_i h_i(\mathbf{x}_j)]^2}{2\sigma^2}}$$

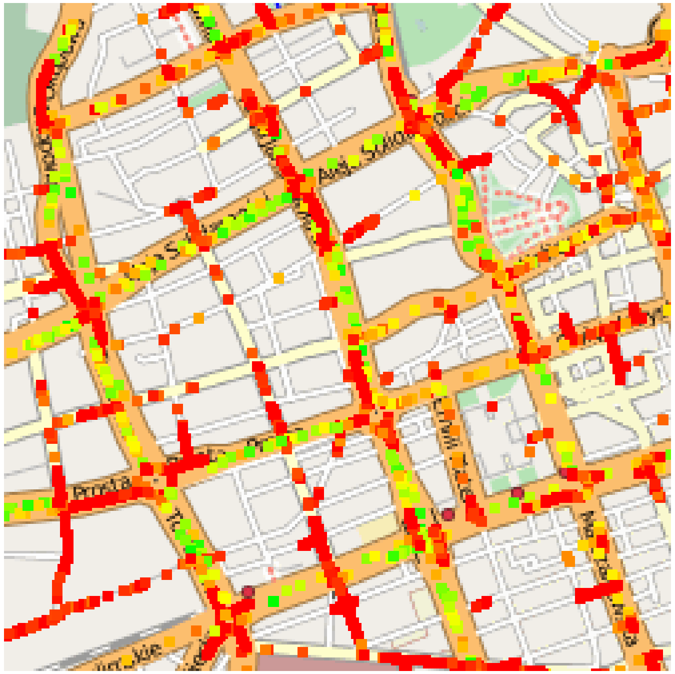
Least-squares Linear Regression is MLE for Gaussians!!!

Applications Corner 1

- Predict stock value over time from
  - past values
  - other relevant vars
    - e.g., weather, demands, etc.

Applications Corner 2

- Predict road traffic volume over time from
  - historical traffic volume
  - historical traffic volume of adjacent road segments



Applications Corner 3

- Predict when a sensor will fail
  - Based on several variables
    - age, chemical exposure, number of hours used,…

- ***Other applications?***

Basics of Linear Algebra

Eigenvector and Eigenvalue

A matrix $A \in \Re^{m \times n}$ is a two-dimensional array

Matrix operations: $A + B$, $A \bullet B$, A^{-1}

$rank(A)$, A^T, $\det(A)$

(λ, x) is an eigen–pair of A, if and only if $Ax = \lambda x$.

λ is the eigenvalue

x is the eigenvector

Orthogonal Matrix

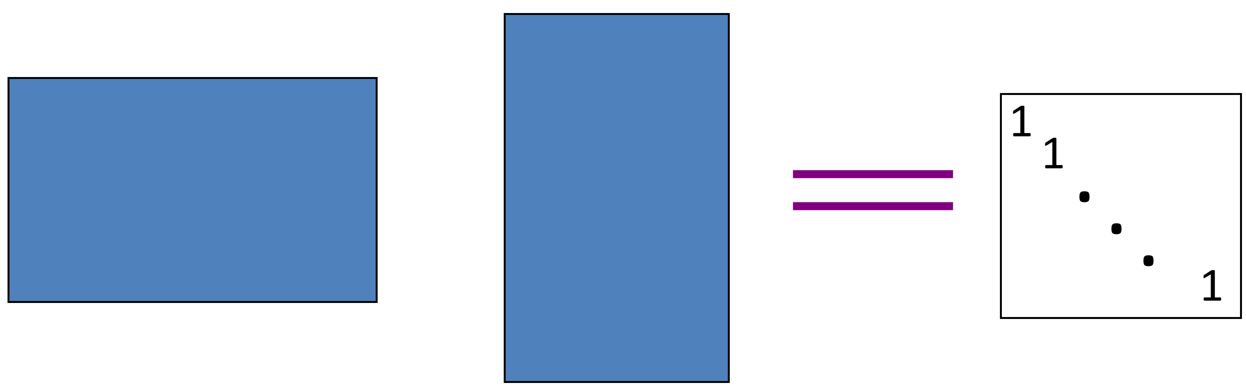
$U \in \Re^{m \times m}$ is orthogonal, if and only if $U\,U^T = I_m$.

(I_m is the identity matrix)

$\Rightarrow \mathrm{U}^{-1} = U^T$

The columns of $V \in \Re^{m \times n} (m > n)$ are orthonormal,

if and only if $V^T V = I$.



Matrix Norms and Trace

Matrix norm :

2 - norm : $\|A\|_2 =$ the square root of the largest eigen value of AA^T.

F - norm : $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^{\ 2}}$.

1 - norm : $\|A\|_1 = \sum_{i,j} |A_{ij}|$.

$\mathrm{trace}(A) = \sum_{i=1}^{m} A_{ii}$, for a square matrix A of size m by m.

$\|A\|_F^2 = \mathrm{trace}(AA^T) = \mathrm{trace}(A^T A)$, $\mathrm{trace}(AB) = \mathrm{trace}(BA)$.

$\|QA\|_F = \|A\|_F$, if Q has orthonormal columns.

Symmetric and Positive Definite Matrix

A is symmetric, if $A = A^T$.

$A \in \Re^{m\times m}$ is symmetric and positive semi-definite,

if $x^T Ax \geq 0$, for any $x \in \Re^m$.

$A \in \Re^{m\times m}$ is symmetric and positive definite,

if $x^T Ax > 0$, for any nonzero $x \in \Re^m$.

If A is symmetric, then all eigenvalues are real.

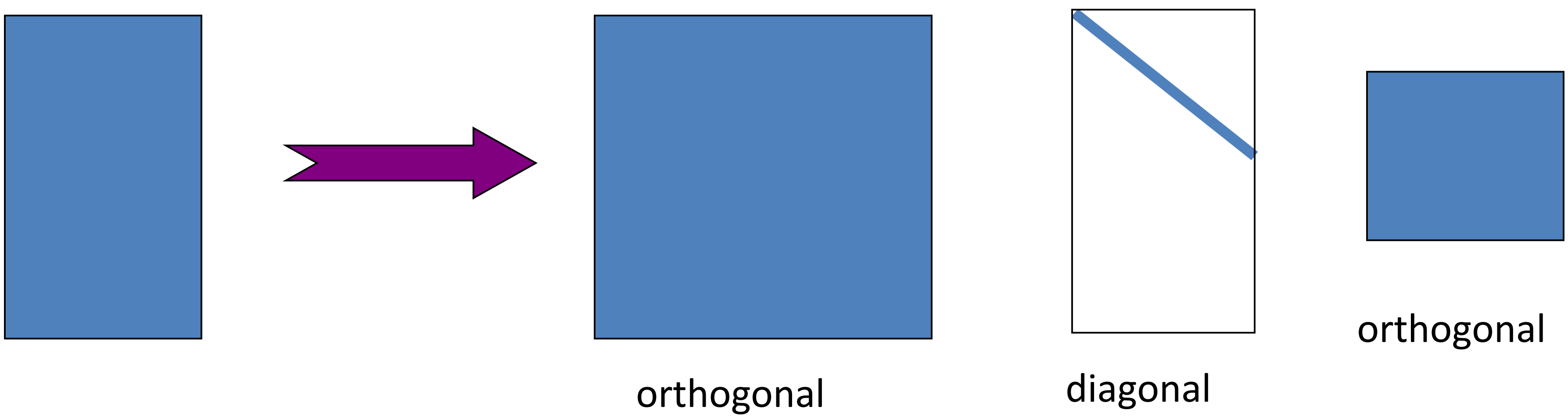
$\Rightarrow A = U\Sigma U^T$, where U is orthogonal and Σ is diagonal.

Singular Value Decomposition

Singular Value Decomposition (SVD): $A = U\Sigma V^T$, where $A \in \Re^{m\times n}$, $U \in \Re^{m\times m}$ and $V \in \Re^{n\times n}$ are orthogonal, and $\Sigma = diag(\sigma_1,\cdots,\sigma_r)$ is diagonal with $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$, and $r = \min(m,n)$.

$AA^T = U\Sigma\Sigma^T U^T$: U forms the eigenvectors of AA^T.

$A^T A = V\Sigma^T\Sigma V^T$: V forms the eigenvectors of $A^T A$.



Some Properties of SVD

THEOREM 2.1. *Let the SVD of A be given by Equation (1) and*

$$\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

and let $R(A)$ and $N(A)$ denote the range and null space of A, respectively. Then,

1. *rank property: $rank(A) = r$, $N(A) \equiv span\{v_{r+1}, \cdots, v_n\}$, and $R(A) \equiv span\{u_1, \cdots, u_r\}$, where $U = [u_1 u_2 \cdots u_m]$ and $V = [v_1 v_2 \cdots v_n]$.*
2. *dyadic decomposition: $A = \sum_{i=1}^{r} u_i \cdot \sigma_i \cdot v_i^T$.*
3. *norms: $\|A\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$, and $\|A\|_2^2 = \sigma_1$.*

Some Properties of SVD

THEOREM 2.2. [Eckart and Young] *Let the SVD of A be given by Equation (1) with* $r = rank(A) \leq p = \min(m, n)$ *and define*

$$A_k = \sum_{i=1}^{k} u_i \cdot \sigma_i \cdot v_i^T \ , \tag{2}$$

then

$$\min_{rank(B)=k} \|A - B\|_F^2 = \|A - A_k\|_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_p^2.$$

- That is, A_k is the optimal approximation in terms of the approximation error measured by the Frobenius norm, among all matrices of rank k
- Forms the basics of LSI (Latent Semantic Indexing) in informational retrieval

Low Rank Approximation by SVD