EECS 445

Introduction to Machine Learning

Regression and Regularization

Prof. Kutty

Linear Regression

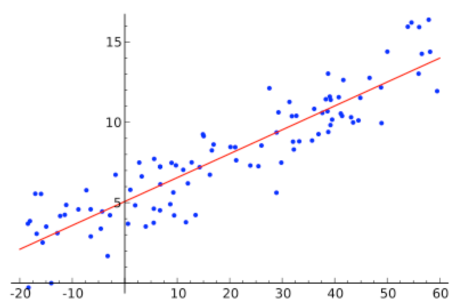
A linear regression function is simply a linear function of the feature vector:

$$f(\bar{x}; \bar{\theta}, b) = \bar{\theta} \cdot \bar{x} + b$$

Learning task:

Choose parameters in response to training set

$$S_n = \{(\bar{x}^{(i)}, y^{(i)})\}_{i=1}^{n} \quad \bar{x} \in \mathbb{R}^d \;\; y \in \mathbb{R}$$



Linear Regression with Squared Loss

$$R_n(\bar{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y^{(i)} - (\bar{\theta}\cdot\bar{x}^{(i)}))^2}{2}$$

SGD with Squared Loss

$k = 0, \bar{\theta}^{(k)} = \bar{0}$

while convergence criteria are not met

randomly shuffle points

for i = 1, …,n

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} + \eta_k (y^{(i)} - \bar{\theta}^{(k)} \cdot \bar{x}^{(i)}) \bar{x}^{(i)}$$

k++

Exact Solution for Regression with Sqd Loss

The parameter value computed as

$$\boxed{\bar{\theta}^* = (X^T X)^{-1} X^T \bar{y}}$$

$$X = [\bar{x}^{(1)}, \ldots, \bar{x}^{(n)}]^T$$

dimension: n x d

exactly minimizes

$$\bar{y} = [y^{(1)}, \ldots, y^{(n)}]^T$$

dimension: n x 1

Empirical Risk with Squared Loss

$$R_n(\bar{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y^{(i)} - (\bar{\theta} \cdot \bar{x}^{(i)}))^2}{2}$$

What if $X^T X$ is singular?

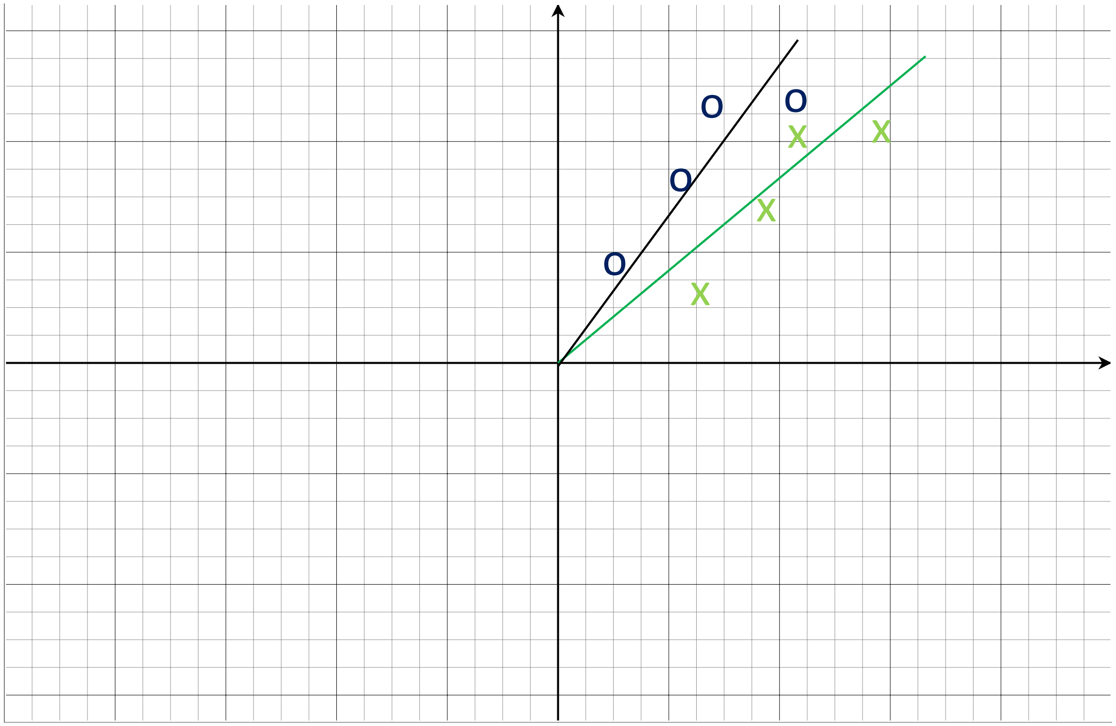
- Why?
  - columns are linearly dependent.
    - implication: features are redundant

- Solution:
  - identify and remove offending features!
  - use **regularization**

$$\boxed{\bar{\theta}^* = (X^T X)^{-1} X^T \bar{y}}$$

Bias-Variance tradeoff

1. Variance



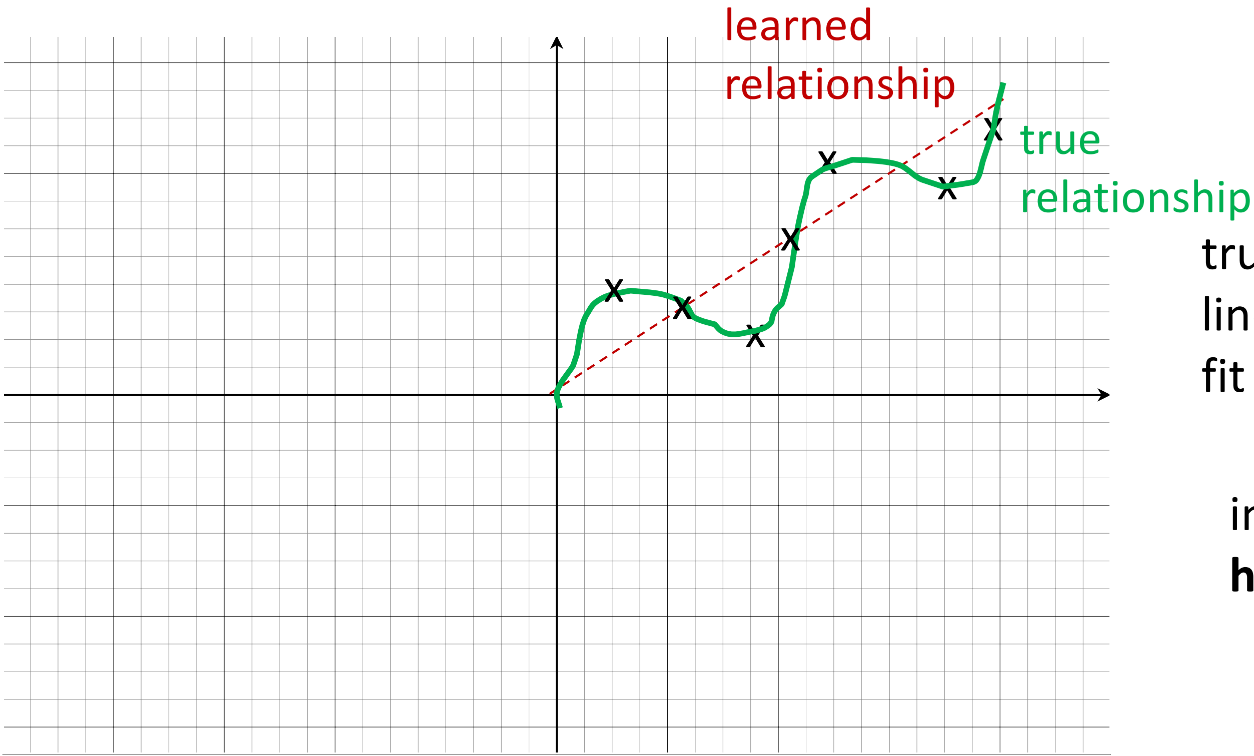
as n increases

variance decreases

Variance is $E_D[\{h(\bar{x};\bar{\theta}) - E_D[h(\bar{x};\bar{\theta})]\}^2]$

measures extent to which the solutions for individual datasets vary around their average

2. Bias



true relationship is non-linear but we are trying to fit data to linear model

increased n does **not help** bias

measures extent to which average prediction over all datasets differs from desired function

Bias^2 is $(E_D[h(\bar{x};\bar{\theta})] - y)^2$

Bias-Variance tradeoff

- to reduce bias, need larger $\mathcal{F}$
- however, if we have noisy/small dataset, this may increase variance
  - Sources of noise:
    - noisy labels
    - noisy features

Bias-Variance Tradeoff

Estimation Error (variance)

*low variance →constant function

*high variance → high degree polynomial, RBF kernel

Structural Error (bias)

*low bias → linear regression applied to linear data, RBF kernel

*high bias → constant function, linear model applied to non-linear data

How to find models that generalize well?

- Feature selection
- Regularization
- Maximum margin separator

As noted earlier, the last two of these are in fact related

Regularization and Ridge Regression

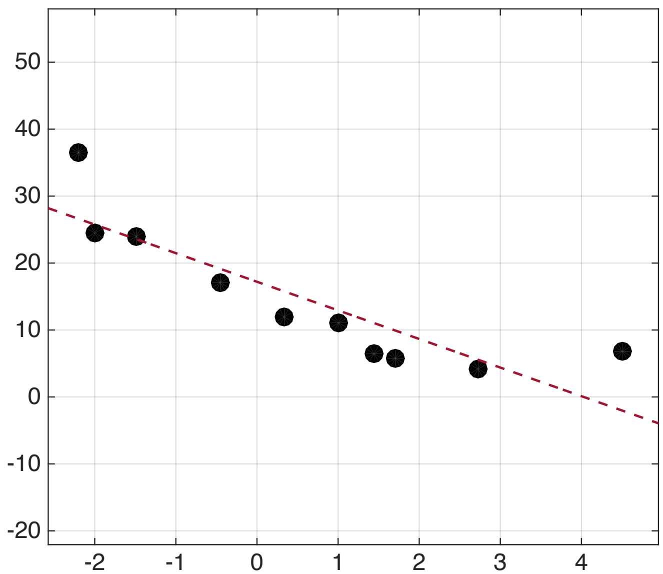
https://forms.gle/ffiBvNbPjHF8ghi77

Regularization

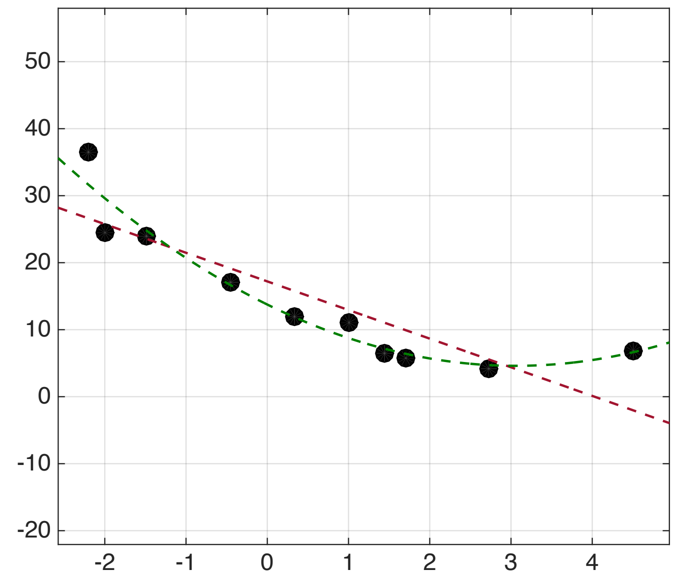
Idea: prefer a *simpler* hypothesis

- will push parameters toward some default value (typically zero)
- resists setting parameters away from default value when data weakly tells us otherwise

Regularization: example



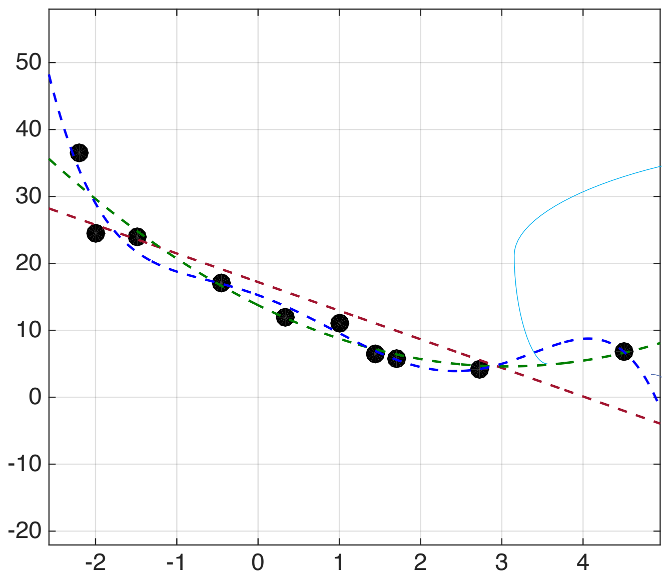
$f(x; \theta, b) = \theta x + b$



$$\phi(x) = [x^2, x, 1]^T$$

$$\bar{\theta} = [1, -6, 13]^T$$

Regularization: example



$$\bar{\theta} = [0, 0, 0, 1, -6, 13]$$

$$\bar{\theta} = [-0.08, 0.5, -0.1133, -4.4, 15]$$

$$\phi(x) = [x^5, x^4, x^3, x^2, x, 1]$$

What should $Z(\bar{\theta})$ be?

- Desirable characteristics:
  - should force components of $\bar{\theta}$ to be small (close to zero)
  - Convex, Smooth
- A popular choice
  - ℓ_p norms
  - Let's use ℓ_2 norm as the penalty function

hyperparameter

$$J_{n,\lambda}(\bar{\theta}) = \boxed{\lambda Z(\bar{\theta})} + R_n(\bar{\theta})$$

regularization term/penalty; λ ≥ 0

Ridge regression

$$J_{n,\lambda}(\bar{\theta}) = \lambda Z(\bar{\theta}) + R_n(\bar{\theta})$$

L2 regularization $$Z(\bar{\theta}) = \frac{||\bar{\theta}||^2}{2}$$

squared loss $$R_n(\bar{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y^{(i)} - (\bar{\theta}\cdot\bar{x}^{(i)}))^2}{2}$$

$$\boxed{J_{n,\lambda}(\bar{\theta}) = \lambda\frac{||\bar{\theta}||^2}{2} + \frac{1}{n}\sum_{i=1}^{n}\frac{(y^{(i)} - (\bar{\theta}\cdot\bar{x}^{(i)}))^2}{2}}$$

Ridge regression

$$J_{n,\lambda}(\bar{\theta}) = \lambda \frac{||\bar{\theta}||^2}{2} + \frac{1}{n} \sum_{i=1}^{n} \frac{(y^{(i)} - (\bar{\theta} \cdot \bar{x}^{(i)}))^2}{2}$$

- when λ = 0
  - this is linear regression with squared loss
- as λ → ∞
  - this is minimized at $\bar{\theta} = \mathbf{0}$
- picking an appropriate λ balances between these two extremes

Ridge regression
Closed form solution

1. Find gradient wrt $\bar{\theta}$
2. Set it to zero and solve for $\bar{\theta}$

$$J_{n,\lambda}(\bar{\theta}) = \lambda \frac{||\bar{\theta}||^2}{2} + \frac{1}{n} \sum_{i=1}^{n} \frac{(y^{(i)} - (\bar{\theta} \cdot \bar{x}^{(i)}))^2}{2}$$

We say $\arg\min_{\bar{\theta}} J_{n,\lambda}(\bar{\theta}) = \bar{\theta}^*$

$$\bar{\theta}^* = (\lambda \mathrm{I} + \mathrm{A})^{-1} \mathrm{b}$$

$$= \boxed{(\lambda' \mathrm{I} + \mathrm{X}^{\mathrm{T}} \mathrm{X})^{-1}} \mathrm{X}^{\mathrm{T}} \bar{y}$$

invertible as long as $\lambda > 0$

Ridge regression
Closed form solution

$$\boxed{\lambda \mathrm{I} + \mathrm{X}^{\mathrm{T}}\mathrm{X}}$$

invertible as long as $\lambda > 0$

Facts:

- A matrix is positive definite iff all its eigenvalues are positive.
- A positive definite matrix is invertible.
- A matrix is positive semi-definite matrix (PSD) iff all its eigenvalues are non-negative.

Claims:

- $X^T X$ is positive semi-definite (PSD).
- If matrix A has eigenvalue k,
  then $A + \lambda I$ has eigenvalue $k + \lambda$.

Soft-Margin SVM: exercise

Claim: Soft margin SVM is an optimization problem with hinge loss as ~~objective~~ the loss function and ℓ_2-norm regularizer

$$\min_{\bar{\theta}, b, \bar{\xi}} \quad \frac{\|\bar{\theta}\|^2}{2} + C \sum_{i=1}^{n} \xi_i \text{ subject to } y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

for $i \in \{1, \ldots, n\}$

Hints:

- Write $\xi_i \geq 1 - y^{(i)}\left(\bar{\theta} \cdot \bar{x}^{(i)} + b\right)$ and $\xi_i \geq 0$
- Observe that the objective function includes the terms $\min_{\bar{\xi}} \sum_{i=1}^{n} \xi_i$

Feature Selection

Feature Selection

Motivation

- When you have few examples and a large number of features (i.e., d>>n) it becomes very easy to overfit your training data
- How can we remove uninformative features?

Different FS Approaches:

① **Filter**

② **Wrapper**

③ **Embedded**

Feature Selection

Filter Approach:

- rank features according to some metric (independent of learning algorithm)
- filter out features that fall below a certain threshold

E.g., correlation with output (i.e., label)

Pearson's correlation

sample means

$$r_{x_j,y} = \frac{\sum_{i=1}^{n}(x_j^{(i)} - \tilde{x}_j)(y^{(i)} - \tilde{y})}{\sqrt{\sum_{i=1}^{n}(x_j^{(i)} - \tilde{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y^{(i)} - \tilde{y})^2}}$$

$$r_{x_{(1)},y}$$

$$r_{x_{(2)},y}$$

$$r_{x_{(3)},y}$$

Threshold

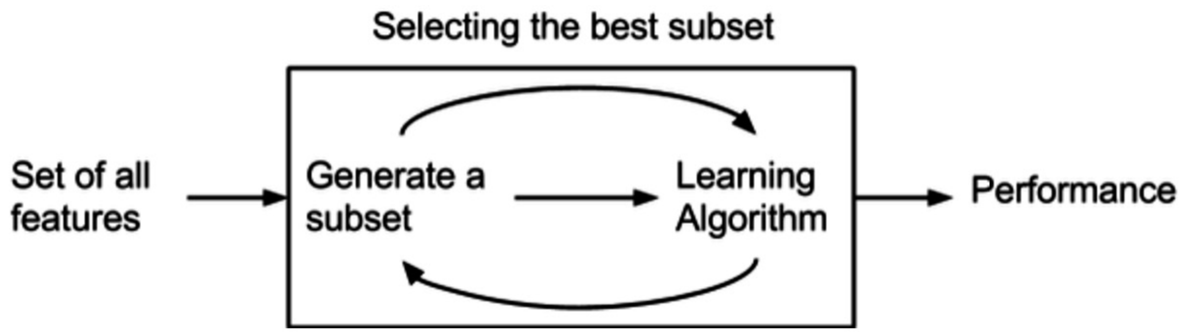
$$\vdots$$

$$r_{x_{(d)},y}$$

Feature Selection

Wrapper Approach:

- utilizes learning algorithm to score subsets according to predictive power
- learning algorithm is “wrapped” in a search algorithm



Feature Selection

	Filter Approach	Wrapper Approach
Pros	• performed only once	• ability to take into account feature dependencies • considers performance of model
Cons	• ignores the performance of the model	• computationally expensive

Feature Selection

Embedded Methods:

- Incorporate variable selection as part of the training process

L2 regularization

$$\min_{\bar{\theta},b,\bar{\xi}} \frac{||\bar{\theta}||^2}{2} + C\sum_{i=1}^{n}\xi_i \quad \text{s.t. } y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)} + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$\text{for } i = 1, \ldots, n$$

L1 regularization

$$\min_{\bar{\theta},b,\bar{\xi}} ||\bar{\theta}||_1 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t. } y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)} + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$\text{for } i = 1, \ldots, n$$

$$\boxed{||\bar{\theta}||_1 = \sum_{i=1}^{d}|\theta_i|}$$

When C is sufficiently small, the L_1-norm penalty will shrink some parameters to *exactly* zero → implicit (or embedded) feature selection end of part 1

Review: Supervised Learning

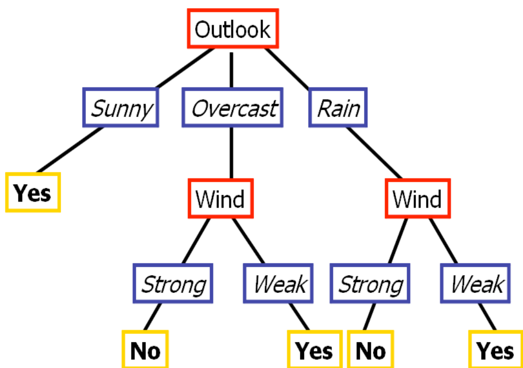
- Perceptron
  - with and without offset
  - convergence
- (Stochastic) Gradient Descent
  - linear classifier with hinge loss
- Support Vector Machines
  - Soft Margin SVMs
  - Kernel trick

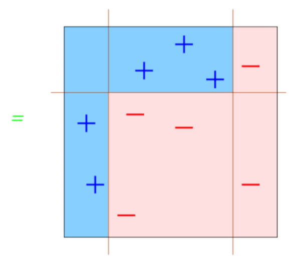
- Regression
  - linear regression with squared loss
    - SGD
    - closed form solution
- Regularization
  - ridge regression
    - SGD
    - closed form solution

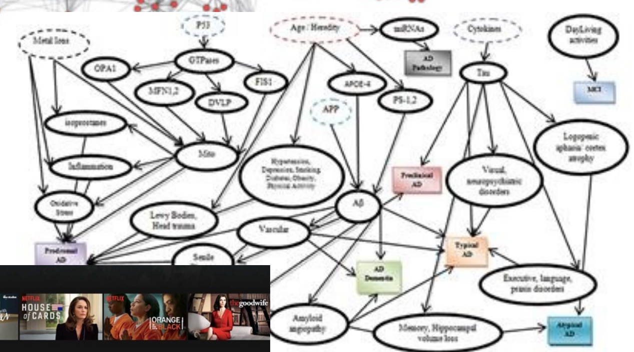
- Neural Networks

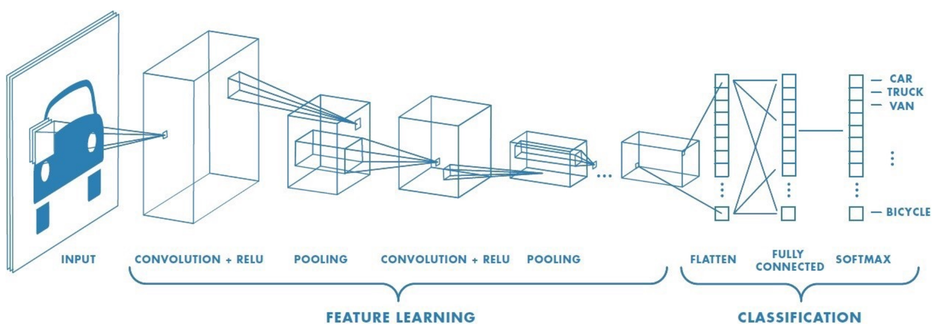
- Decision trees
- Boosting
- Ensemble Methods

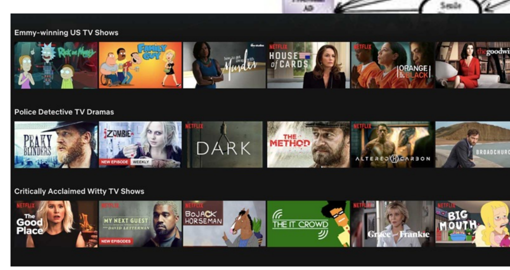
Coming up in parts 2 and 3

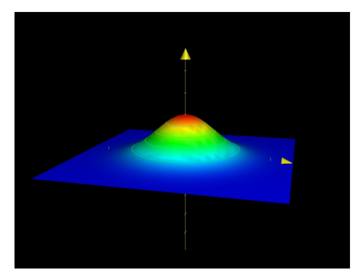












Breaking news…

I'm offering a new course in Fall 2024:

Machine Learning Research Experience

Are you **curious about research** and looking for an opportunity to try it? Have you worked in a research lab but are looking for further autonomy and the ability to **propose new ideas**? Are you interested in taking an in-depth look at **cutting-edge Machine Learning research** and testing them out yourself? If so, this course might be for you!

Course details* will be provided here so watch that space!

*can count as MDE/Capstone for CS/CE majors

CSE Values

HACKS

Honesty
Conduct ourselves with integrity and communicate with transparency and authenticity.

Achievement
Strive for academic excellence and celebrate personal and collective efforts and accomplishments.

https://forms.gle/ffiBvNbPjHF8ghi77

Cooperation
Collaborate in work and learning, promote inclusion and mutual respect, encourage diverse perspectives, and look after each other.

Knowledge
Protect academic freedom, advance learning and scientific progress, and cultivate wisdom.

Service
Contribute to the well-being of our community and global society.

so long… for now