EECS 445

Introduction to Machine Learning

Regression

Prof. Kutty

Today's Agenda

- Project 1 due Tuesday 2/13 at 10pm. Please remember to upload appropriate components to Gradescope.

- Deadline for exam conflicts has passed (both midterm and final)
  - we are aware of conflicts with STATS250 (final)
  - EECS 376/EECS 492 (midterm)
  - SSD accommodations start at 5pm for the final exam
  - More details about midterm/final will be available closer to the exam dates

Review: SVM

$S_n = \{(\bar{x}^{(i)}, y^{(i)})\}_{i=1}^n$

$\downarrow$

$S_n' = \{(\phi(\bar{x}^{(i)}), y^{(i)})\}_{i=1}^n$

Support Vector Machines

Quadratic Program formulation

$$\min_{\bar{\theta}} \frac{\|\bar{\theta}\|^2}{2} \text{ subject to } y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) \geq 1 \text{ for } i \in \{1, \dots, n\}$$

original problem: $\min_{\bar{w}} f(\bar{w})$ s.t. $h_i(\bar{w}) \leq 0$ for $i = 1, \dots, n$

Lagrangian: $L(\bar{w}, \bar{\alpha}) = f(\bar{w}) + \sum_{i=1}^n \alpha_i h_i(\bar{w})$ $\alpha_i \geq 0$

1. Compose the Lagrangian

   $$L(\bar{\theta}, \bar{\alpha}) = \frac{\|\bar{\theta}\|^2}{2} + \sum_{i=1}^n \alpha_i \left(1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})\right) \text{ with } \alpha_i \geq 0$$

2. Write the dual formulation

   $$\max_{\bar{\alpha}, \alpha_i \geq 0} \min_{\bar{\theta}} L(\bar{\theta}, \bar{\alpha})$$

3. Rewrite in primal variable in terms of dual variables

   Set $\nabla_{\bar{\theta}} L(\bar{\theta}, \bar{\alpha}) |_{\bar{\theta} = \bar{\theta}^*} = 0 \rightarrow \bar{\theta}^* = \sum_{i=1}^n \alpha_i y^{(i)} \bar{x}^{(i)}$

4. Simplify the dual formulation

Dual formulation

$$\max_{\bar{\alpha}, \alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)} \cdot \bar{x}^{(j)}$$

Kernelized Dual SVM

$K(\bar{x}^{(i)}, \bar{x}^{(j)}) = \phi(\bar{x}^{(i)}) \cdot \phi(\bar{x}^{(j)})$

$$\max_{\bar{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \boxed{(\phi(\bar{x}^{(i)}) \cdot \phi(\bar{x}^{(j)}))}$$

$$\text{subject to } \alpha_i \geq 0 \quad \forall i = 1, .., n$$

$$\max_{\bar{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \boxed{K(\bar{x}^{(i)}, \bar{x}^{(j)})}$$

$$\text{subject to } \alpha_i \geq 0 \quad \forall i = 1, .., n$$

- Sometimes it is *much more* efficient to compute $K(\bar{x}^{(i)}, \bar{x}^{(j)})$ directly
- Intuitively, can think of $K(\bar{x}^{(i)}, \bar{x}^{(j)})$ as a measure of similarity between $\bar{x}^{(i)}$ and $\bar{x}^{(j)}$

Examples of Valid Kernels

Linear Kernel

$$K(\bar{u}, \bar{v}) = \bar{u} \cdot \bar{v}$$

Quadratic Kernel

$$K(\bar{u}, \bar{v}) = (\bar{u} \cdot \bar{v} + r)^2 \text{ with } r \geq 0$$

RBF Kernel (aka Gaussian Kernel)

$$K(\bar{u}, \bar{v}) = \exp(-\gamma \|\bar{u} - \bar{v}\|^2) \quad \text{with } \gamma \geq 0$$

Kernel algebra

Let K1 and K2 be valid kernels, then the following are valid kernels:

$$K(\bar{x}, \bar{z}) = K_1(\bar{x}, \bar{z}) + K_2(\bar{x}, \bar{z}) \quad \text{sum}$$

$$K(\bar{x}, \bar{z}) = \alpha K_1(\bar{x}, \bar{z}) \quad \text{scalar product } \alpha > 0$$

$$K(\bar{x}, \bar{z}) = K_1(\bar{x}, \bar{z}) K_2(\bar{x}, \bar{z}) \quad \text{direct product}$$

RBF Kernel Feature Map

(Proof Sketch)

$K(\bar{u}, \bar{v}) = \exp(-\gamma \|\bar{u} - \bar{v}\|^2)$

$K(\bar{u}, \bar{v}) = \exp(-\gamma \|\bar{u} - \bar{v}\|^2)$

$\bar{u} \in \mathbb{R}^2$

$\bar{v} \in \mathbb{R}^2$

$\|\bar{u} - \bar{v}\|^2 = \|(u_1 \;\; u_2)^T - (v_1 \;\; v_2)^T\|^2$

$= \left\| \begin{bmatrix} u_1 - v_1 \\ u_2 - v_2 \end{bmatrix} \right\|^2$

$= (u_1 - v_1)^2 + (u_2 - v_2)^2$

$= u_1^2 + v_1^2 - 2u_1v_1 + u_2^2 + v_2^2 - 2u_2v_2$

$= \|\bar{u}\|^2 + \|\bar{v}\|^2 - 2(\bar{u} \cdot \bar{v})$

$K(\bar{u}, \bar{v}) = \exp(-\gamma \|\bar{u}\|^2 - \gamma \|\bar{v}\|^2 + 2\gamma(\bar{u} \cdot \bar{v}))$

$= \exp(-\gamma \|\bar{u}\|^2) \exp(-\gamma \|\bar{v}\|^2) \exp(2\gamma(\bar{u} \cdot \bar{v}))$

Recall Taylor series expansion

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Idea $$e^{\bar{u}\cdot\bar{v}} = \frac{(\bar{u}\cdot\bar{v})^0}{0!} + \frac{(\bar{u}\cdot\bar{v})^1}{1!} + \frac{(\bar{u}\cdot\bar{v})^2}{2!} + \cdots$$

Sum of polynomial kernels

$\longrightarrow$ feature map of the RBF kernel is infinite dimensional

Mercer's Theorem

Intuition

A function $K: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a valid kernel

iff

for any $\overline{\boldsymbol{x}}^{(1)}, \dots, \overline{\boldsymbol{x}}^{(n)}$ with $\overline{\boldsymbol{x}}^{(i)} \in \mathbb{R}^d$ and finite n

the $\boxed{n \times n \text{ matrix G with } G_{ij} = K(\overline{\boldsymbol{x}}^{(i)}, \overline{\boldsymbol{x}}^{(j)})}$

is **positive-semidefinite**

That is, G is

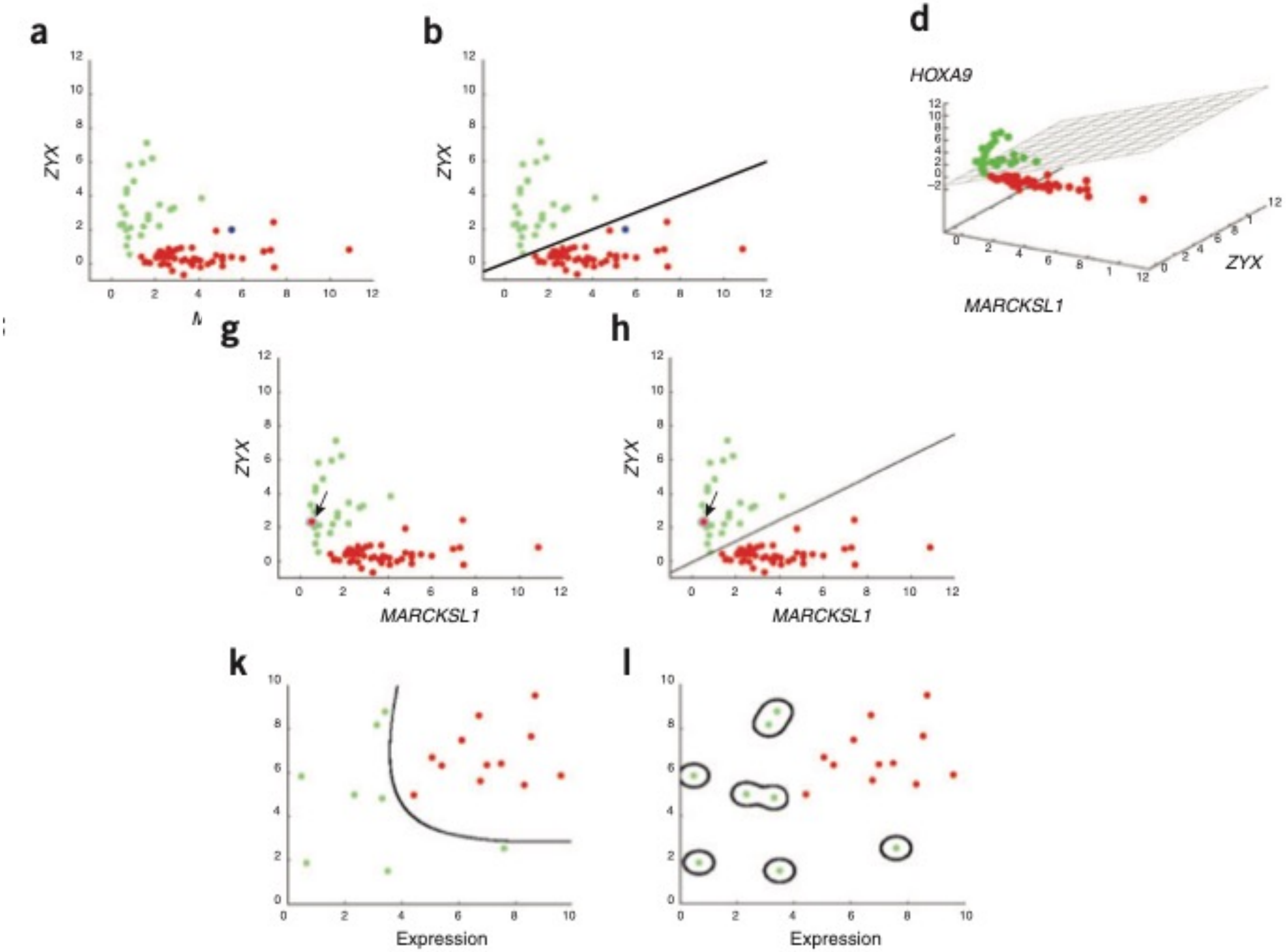
- symmetric $G = G^T$
- and $\forall \bar{z} \in \mathbb{R}^n \quad \bar{z}^T G\, \bar{z} \geq 0$

In other words

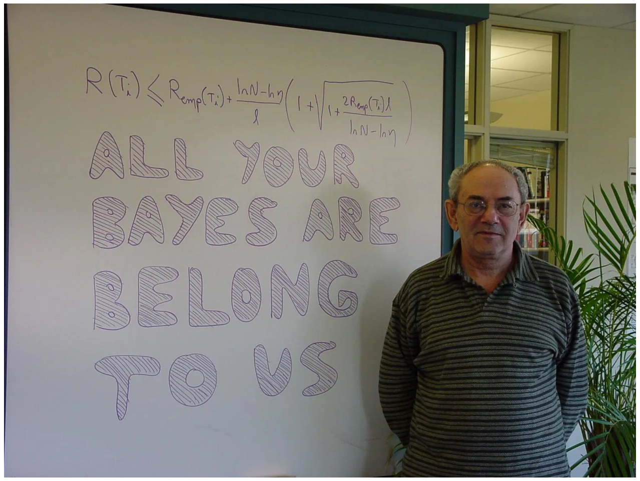
for such a function $K(\bar{u}, \bar{v})$, there exists a function φ such that $K(\bar{u}, \bar{v}) = \varphi(\bar{u}) \cdot \varphi(\bar{v})$

Support vector machines (SVMs) at work:

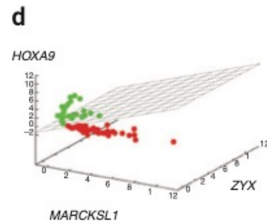
distinguishing acute lymphoblastic leukemia from acute myeloid leukemia (AML).



What is a support vector machine? William S Noble (NATURE BIOTECHNOLOGY 2006)



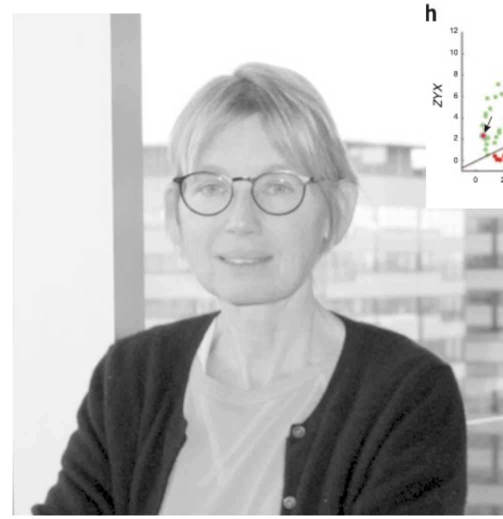
Vladimir Vapnik

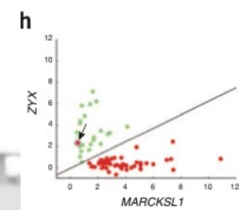


Alexey Chervonenkis

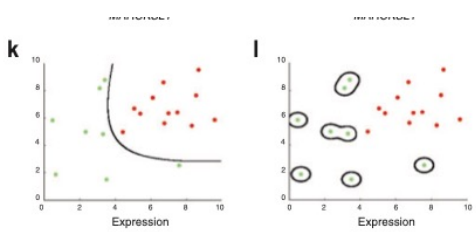
Isabelle Guyon

Bernhard Schölkopf





Corinna Cortes

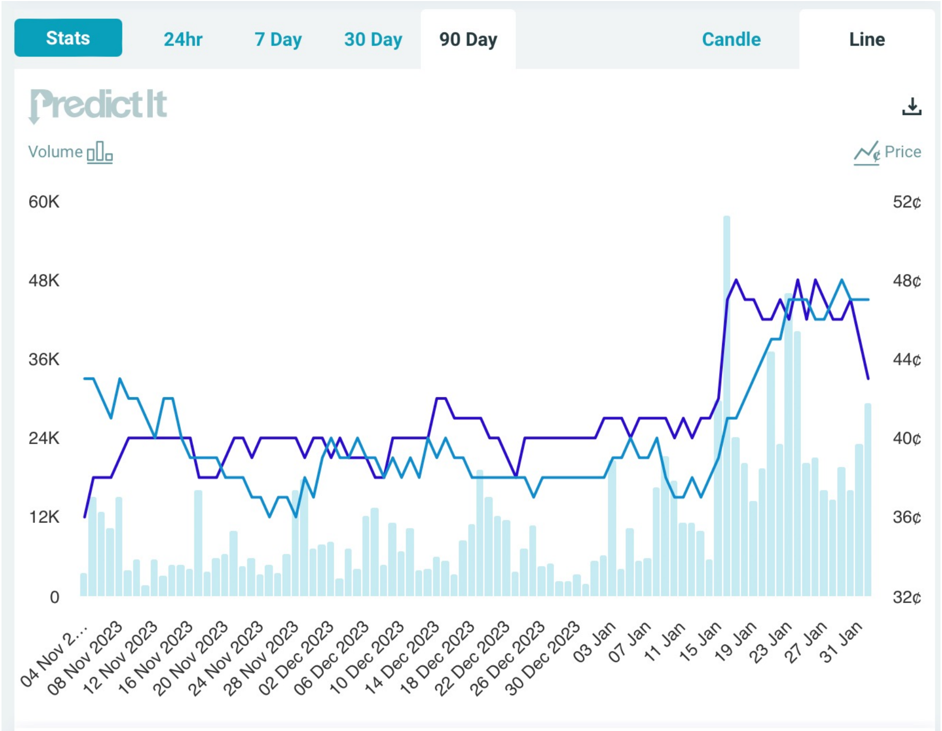


Regression

Supervised Learning

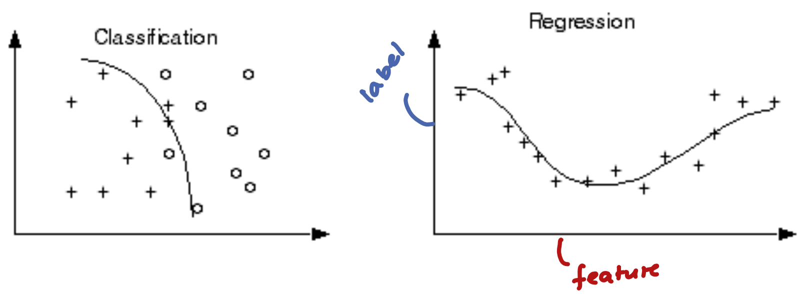
- Goal:
  - Given data $\mathcal{X}$ (in feature space) and the labels $\mathcal{Y}$
  - Learn to predict $\mathcal{Y}$ from $\mathcal{X}$
- Labels could be discrete or continuous
  - Discrete labels: **classification**
  - Continuous labels: **regression**

e.g., 2024 US Presidential Election
Vote Share Market



Regression vs Classification

Classification problem: $y \in \{-1, 1\}$ or $y \in \{0, 1\}$



Regression problem: $y \in \mathbb{R}$

Regression function $f : \mathbb{R}^d \to \mathbb{R}$ where $f \in \mathcal{F}$

Linear Regression

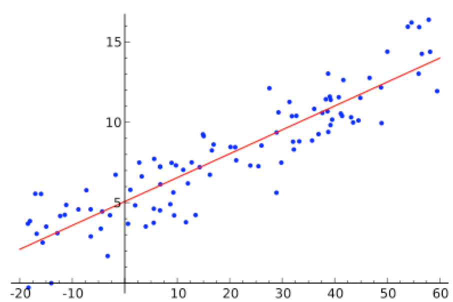
A linear regression function is simply a linear function of the feature vector:

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

Learning task:

Choose parameters in response to training set

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



Empirical risk for Linear Regression

Given $\{(\bar{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$

Recall empirical risk for linear classification

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

true label

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

prediction

Linear Regression with Squared Loss

Least Squares Loss function

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

Squared Loss:

$$\text{Loss}(z) = \frac{z^2}{2}$$

Idea:

permit small discrepancies

heavily penalize large deviations

Empirical risk with sqd 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

Reminder: Squared loss $\text{Loss}(z) = \frac{z^2}{2}$

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

while convergence criteria are not met

randomly shuffle points

for i = 1, …,n

$$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} - \eta_k \boxed{\nabla_{\bar{\theta}} \text{Loss}_{sqd}\left(y^{(i)} - \left(\bar{\theta} \cdot \bar{x}^{(i)}\right)\right)}$$

k++

$$\nabla_{\bar{\theta}} \left(\frac{\left(y^{(i)} - \left(\bar{\theta} \cdot \bar{x}^{(i)}\right)\right)^2}{2} \right)$$

Least Squares Loss function

$$\nabla_{\overline{\theta}} \; \frac{(y^{(i)} - (\bar{\theta} \cdot \bar{x}^{(i)}))^2}{2}$$

$$= \frac{1}{2} \cdot 2 \left(y^{(i)} - \bar{\theta} \cdot \bar{x}^{(i)}\right) \left(- \bar{x}^{(i)}\right)$$

$$= - \left(y^{(i)} - \bar{\theta} \cdot \bar{x}^{(i)}\right) \bar{x}^{(i)}$$

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++

Closed form solution for Empirical Risk with Squared Loss

Optimal value of $\bar{\theta}$ for $R_n(\bar{\theta})$

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

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

Find gradient, set to 0 and solve for $\bar{\theta}$

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

$-\bar{b} + A\,\bar{\theta}^* = 0$

$$\nabla_{\bar{\theta}} R_n(\bar{\theta})|_{\bar{\theta}=\bar{\theta}^*} = 0$$

$$= -\underbrace{\boxed{\frac{1}{n}\sum_{i=1}^{n} \bar{x}^{(i)} y^{(i)}}}_{\bar{b}} + \underbrace{\boxed{\frac{1}{n}\sum_{i=1}^{n} \bar{x}^{(i)} (\bar{x}^{(i)})^T}}_{A}\,\bar{\theta}^*$$

$\bar{b}$ dimension: d x 1

A dimension: d x d

$$\bar{\theta}^* = A^{-1}\,\bar{b}$$

Alternative notation

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n}\bar{x}^{(i)}y^{(i)} = \frac{1}{n}X^T\bar{y}$$

$$A = \frac{1}{n}\sum_{i=1}^{n}\bar{x}^{(i)}(\bar{x}^{(i)})^T$$

$$= \frac{1}{n}X^TX$$

convince yourself of this!

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

$$= \begin{bmatrix} x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \cdots & x_d^{(n)} \end{bmatrix}$$

dimension: n x d

$$\bar{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

dimension: n x 1

Exact Solution for Regression

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

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

If an exact solution exists, why use SGD?

short answer*: efficiency*

What if X^TX is singular?

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

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

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