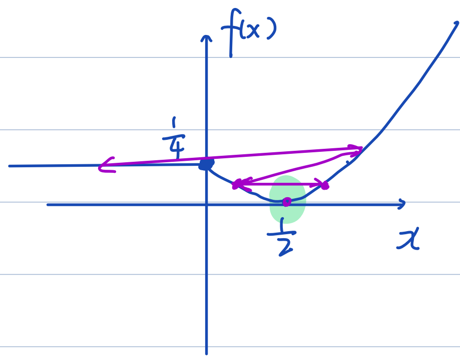
1. Devise the update rule for minimizing $f(x) = (\max(0, x) - \frac{1}{2})^2$ using gradient descent. Roughly, when will gradient descent succeed or fail? (based on where you start and step size)

① Draw the plot for $f(x) = (\max(0, x) - \frac{1}{2})^2$

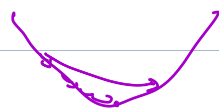
$$= \begin{cases} \frac{1}{4}, & x \le 0 \\ (x - \frac{1}{2})^2, & x > 0 \end{cases}$$

Continuous. $\frac{1}{4} = (0 - \frac{1}{2})^2$



② Compute

$$f'(x) = \begin{cases} 0, & x \le 0 \\ 2(x - \frac{1}{2}) = 2x - 1, & x > 0 \end{cases}$$



③ $x_0 \le 0$ will fail because it always get 0 gradient.

④ $x_{n+1} = x_n - \eta \cdot \nabla$

$x_0 > 0$ $\quad x_{n+1} = x_n - \eta \cdot (2x - 1)$.

Step size need to be reasonablely small. Otherwise, it will diverge.

Let's say $x_0 = \frac{1}{4}$ $\quad \eta = 1$ $\quad x_1 = \frac{1}{4} - 1 \cdot (-\frac{1}{2}) = \frac{3}{4}$,

$x_2 = \frac{3}{4} - 1 \cdot (\frac{3}{4} \times 2 - 1) = \frac{1}{4}$

x will be trapped in $\frac{1}{4}$ & $\frac{3}{4}$.

$\eta = 0.01$

$\times$ in lecture

2. Given a data set A, under what circumstances will its projection onto its principal components equal its projection on

   1. Right singular vectors A $m > n$
   2. Left singular vectors. A $n > m$

In PCA, we find the principal components by calculating the eigendecomposition of A^TA or AA^T given a data set A.

If A is $\underset{\text{sample}}{m} \times \underset{\text{features}}{n}$, $\underset{n \times n}{A^TA} \longrightarrow$ sample covariance matrix.

$C_{i,j}$: How similar feature i to feature j)

If A is $n \times m$. $\underset{n \times n}{AA^T}$.

Q

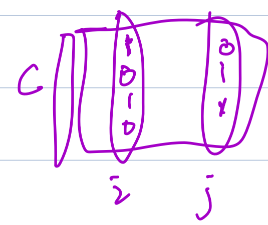
Data $A = \underset{\text{left}}{U} S \underset{\text{right}}{V^T}$ $A^TA = VS(U^TU)SV^T = VS^2V^T$ right singular vector

$AA^T = US(V^TV)SU^T = US^2U^T$ Left ....

$i \in T$ $j \in S$

3. Let S be a set of documents and let T be set of terms. Suppose that C is a binary term-document incidence matrix. (So entry $<i,j>$ is 1 if term i appear in document j and 0 otherwise).

What do the entries of C^TC represent?



1+

① Dimension of C: # term $\times$ # docs.

② Dimension of C^TC: # docs $\times$ # docs

$C^TC<i,j>$ = ith row of C^T · jth column of C

= ith column of C · jth column of C.

column k of C means if each term is in doc k.

= # shared terms in doc i & j.

If $i=j$, # terms in doc i/j.

4. How did we derive the equations for simple linear regression from class.

Page 19 - 20 from CS 378H.

Single variable $\beta_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$

Multiple - Variable Case.

X $m \times n$. $y \in R^m$. Goal: minimize $\| X \cdot w - y \|^2$.

Normal Equation $w = (X^T X)^{-1} X^T y$. $O(n^3 + mn^2)$

Take gradient to $\underline{\| X \cdot w - y \|_2^2}$ with respect to $\underline{w}$.

$\nabla = 2(\underbrace{\underbrace{X \cdot w}_{m\times n \cdot n\times 1} - y}_{m \times 1})^T \cdot \underbrace{X}_{m \times n} = 0$

$2 \, \underline{(X \cdot w)^T \cdot X} = 2 \, \underline{y^T \cdot X}$

$\underline{(X^T X)^{-1} X^T} X w = X^T y$

$w = \underline{(X^T X)^{-1}} X^T y$.

5. Why can you assume without loss of generality that a mistake - bounded Learner only updates its state when it makes a mistake ?

Re-ordering the samples

No matter you update the state or not when you predict it right. It always need to make $\underline{t \text{ mistakes}}$ to ensure no mistakes in future. So updating its state when predicting right is redundant.

6. You are given a data set with m points and an algorithm that satisfies the weak-learing condition. (It always outputs a classifier with accuracy 60%). Each classifier output by the weak-learning algorithm can be encoded using two bits. How can you construct a classifier that can be described by less than m bits and is correct on every data point in the data set. (You may assume m is very large). Trainning.

What's the size of your final classifier?

T

$\varepsilon \leq \frac{1}{m} \to 0$

Adaboost.

After T iterations of the algorithm, the error of h_{final} = MAJ $(h_1, \cdots h_T)$ is at most $e^{-2T\gamma^2}$.

Majority

$T \geq \frac{\log(\frac{1}{\varepsilon})}{2\gamma^2}$

makes the error of h_{final} at most ε. γ, ε

Since m is very large all correct means

$\varepsilon \leq \frac{1}{m}$

$\gamma = 1 - \varepsilon \approx 1$

$T \geq \log(m)$ $O(\log(m))$

$\gamma = 0.1$

$$T \geq \frac{\log(\frac{1}{\varepsilon})}{\gamma^2}$$

$$= \frac{\log(m)}{\gamma^2}$$

$$= \log(m)$$

$\varepsilon = \frac{1}{m}$

$\frac{1}{\varepsilon} = m$

$\gamma^2 = 1 - \varepsilon = 1$

7. Markov's inequality. [weakest]

$\forall X \geq 0$, $P(X \geq a) \leq \dfrac{E(X)}{a}$

non-negative $a > 0$

$E(X) = \cancel{P(X<a) \cdot E(X \mid X<a)} + P(X \geq a) \cdot E(X \mid X \geq a)$

$E(X) \geq P(X \geq a) \cdot \underbrace{E(X \mid X \geq a)}_{\geq a.} \geq \underline{a \cdot P(X \geq a)}$

Chebyshev's Inequality : (X random variable with mean u and variance σ^2)

$P_r(|X-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}$ for any $k > 0$.

$P_r(|X-\mu| \geq k\sigma) = \underbrace{P_r((X-\mu)^2 \geq k^2\sigma^2)}_{\text{Apply Markov inequality}} \leq \dfrac{E[(X-\mu)^2]}{k^2\sigma^2} = \dfrac{\sigma^2}{k^2\sigma^2} = \dfrac{1}{k^2}$

] $Y = (X-\mu)^2$ with $a = k^2\sigma^2$

$X = (x-\mu)^2 \qquad a = k^2 a^2.$

Markov : $P(X \geq a) \leq \dfrac{E[(X-\mu)^2]}{k^2 a^2} = \dfrac{\sigma^2}{k^2\sigma^2} = \dfrac{1}{k^2}$

Chernoff Bounds (Strongest)

$P_r(X \geq a) = P_r(e^{t \cdot X} \geq e^{t \cdot a}) \leq \dfrac{E(e^{t \cdot X})}{e^{t \cdot a}}$