Unit 2 Classifiers and K means

Sunday, February 26, 2023 5:40 PM

Additional notes

$(AB)^{-1} = B^{-1}A^{-1}$

$(AB)^T = B^T A^T$

- what is a classifier and how do you make one?

it's a way of catigorizing data. you create one by using training data (X) and labels (Y) in this way you can solve the least squares problem for the weights vector w often times we use a decision rule to make clear the boundaries of our classifications

for ex) $\tilde{y} = \begin{cases} -1 & \text{if } w_1 \text{length} + w_2 \cdot \text{lightness} > t \\ 1 & \text{otherwise} \end{cases}$

X would equal $= \begin{bmatrix} \text{length} & \text{lightness} & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}$

$w^T = [w_1\ w_2\ w_3]$ where $w_3 = -t$

$y = [y_1, y_2, y_3 \cdots y_n]^T$

If the cols of X are lin indep, we can find $\hat{w} = (X^TX)^{-1}X^Ty$

- how do you write a system of linear eqs in matrix form?

$\begin{matrix} 5x + 6y - 2 = 0 \\ 7x + 2 = 5 \end{matrix} \Rightarrow \underset{2\times3}{\begin{bmatrix} 5 & 6 & -2 \\ 7 & 0 & 2 \end{bmatrix}} \underset{3\times1}{\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}} = \underset{2\times1}{\begin{bmatrix} 0 \\ 5 \end{bmatrix}}$

- what is lin indep?
  - how to tell if 2 vectors are lin indep?
  - what can lin indep tell us about a problem?

2 vectors are lin independent if they are not multiples of eachother

3+ vectors are lin independent if $V_1\alpha_1 + V_2\alpha_2 \neq V_3$ (side note, orthogonal vectors are alway lin indep but just because 2 vectors are lin indep does not mean they are orthogonal. to prove orthogonality take the dot product of 2 vectors, if it's zero, the vectors are orthogonal. ex) $[1\ 1] \cdot [-1\ 1] = -1+1 = 0$ ✓)

- how do we use rank? what is "the rank of a matrix"

rank is the # of lin indep rows/cols of a matrix

1) $\text{rank}\{A\} < \text{rank}\{[A + d]\}$ = No sol

2) $\text{rank}\{A\} = \text{rank}\{A+d\}$ & $\text{rank}\{A\} = \dim\{w\}$ = 1 soln

3) $\text{rank}\{A\} = \text{rank}\{A+d\}$ & $\text{rank}\{A\} < \dim\{w\}$ = ∞ soln

- All full rank (P) square matrices are invertable.

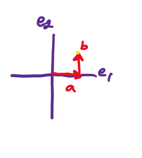
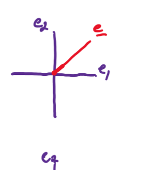
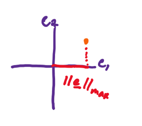
A^TA is invertable if $\underset{N\times P}{A^T}\ \underset{P\times N}{A}$ is rank P

- what is a norm? how can you tell if a function $f(x)$ is a valid norm? what is the purpose of a norm?

  - a vector norm measures the "size" of a vector

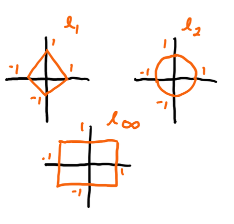
A vector norm $\|\cdot\|$ map from $\mathbb{R}^n \to \mathbb{R}$ and satisfies the following

1) $\|x\| \geq 0$ for all x
2) $\|x\| = 0$ only if $x = 0$
3) $\|bx\| = |b|\|x\|$ for all $b \in \mathbb{R}$ and $x \in \mathbb{R}$
4) Triangle Inequality $\|x+y\| \leq \|x\| + \|y\|$

- $\|e\|_1 = \sum_i |e_i| = a+b$  "Taxi-cab norm" ℓ_1
- $\|e\|_2 = (\sum_i |e_i|^2)^{1/2}$  "euclidean norm" ℓ_2
- $\|e\|_\infty = \max_i |e_i|$  "max norm" ℓ_∞
- $\|e\|_\ell = (\sum_i |e_i|^\ell)^{1/\ell}$ ℓ_ℓ norm

unit ball: $\{x : \|x\| = 1\}$

every point on the boundary has unit 1



"The functions look a little like $F(x) = 1 = \|x\|_1 + \|x\|_\infty$"

$(A^TA)^{-1} = I$

if it's orthonormal not necessarily

Side note: in 3D $\hat{d} - d \perp \text{span}\{a_1, a_2\}$ plane

- least squares solution $Aw = d$ (assume $N \geq P$, rank$(A) = P$)

Goal: to minimize $\|Aw - d\|_2^2$ d lies in p-dim subspace and is spanned by cols of A

it stems from an

$\boxed{\text{in general } A^T(\hat{d} - d) = 0}$ "orthogonality condition"

"when they ask you to find the value of the squared error $\min_w \|Xw - y\|_2^2$, they give you X and y, find $\hat{w}$ and plug it back in"

$A^T(Aw - d) = 0$

$A^TAw = A^Td$

$(A^TA)^{-1}A^TAw = (A^TA)^{-1}A^Td$

$w = (A^TA)^{-1}A^Td$

$w = (A^TA)^{-1}A^Td$

You can also find w by taking a gradient

$\|Aw - d\|_2^2 = (Aw-d)^T(Aw-d) = w^TA^TAw - w^TA^Td - d^TAw + d^Td \Rightarrow f(w)$ quadratic in w

To minimize w you take the derivative and set it equal to zero

Gradient = derivative by a vector

$\nabla_w f(w) = \left[\frac{\partial}{\partial w_1} f(w) + \cdots \frac{\partial f(w)}{\partial w_p}\right]^T$

Suppose $f(w) = w^TQw$, then $\nabla_w f(w) = Q^Tw + Qw$ (why?)

if we look at the symmetric case where $Q^T = Q$, then

$\nabla_w f(w) = 2Qw$

So, sol $\nabla_w(w^TA^TAw - w^TA^Td - d^TAw + d^Td) = 0$

$2A^TAw - A^Td - A^Td = 0$

$\boxed{w = (A^TA)^{-1}A^Td}$

provide an example of taking the gradient of linear and quad functions of a vector

ask about hw 4.

exam 2 prep

exam 1 sol

how would you recommend I study

what more should I know: can you ask me a question?

- what are positive definite and positive semidefinite matrices and what do we use them for?

  - what is it equivalent to rank full => positive definite sol

- how do you project vectors onto a subspace? if it's not diagonal, perform a basis transformation

Projection

- $\hat{d} = A(A^TA)^{-1}A^Td = P_A d$ = Projecting the vector d onto the span of the cols of A
- $P_A = A(A^TA)^{-1}A^T$ = Projection matrix
- $P_A^2 = P_A$
- $P_{A\perp} = I - P_A$ this projects onto the space $\perp$ to span$\{A\}$

The span A = the rank A

- GS process
  1. set $\tilde{u}_1 = x_1$
  2. normalize $u_1 = \frac{\tilde{u}_1}{\|\tilde{u}_1\|_2}$
  3. set $\tilde{u}_2 = x_2 - \text{proj}_{u_1} x_2$
  4. normalize $u_2 = \frac{\tilde{u}_2}{\|\tilde{u}_2\|_2}$
  5. $\tilde{u}_3 = x_3 - \text{proj}_{u_1} x_3 - \text{proj}_{u_2} x_3$

$\text{Proj}_u a = u(u^Ta)$

2-D.R

Tell me about the construct of GS + etc

- Orthonormal basis, unit 1 and orthogonal
- if your training data is not full rank

Properties of orthonormal basis

1. $m \leq n$ (dimension of subspace) $m = n$ if $\bar{x}_1 \ldots \bar{x}_n$ is lin indep
2. $m \leq \dim(X_i)$
3. $U^TU = I_{m\times m}$
4. U is square, then $U^{-1} = U^T$
5. if U is square, $UU^T = I_{m\times m}$

$(A^TA) = I$ if your full rank ie orthonormal basis

- what's the benefit of designing classifiers using orthonormal bases (are we talking about X being orthonormal?)

- what is the Tikhonov-regularized least squares problem
  - penalizes sols with large norm $\|w\|^2$

$w^* = \min_w \|Aw - d\|^2 + \lambda\|w\|^2$

$w = (A^TA + \lambda I)^{-1}A^Td$

- always unique always exists

small λ gets bias small but increase variance (ability to predict training sample; sensitivity to train)

large λ increases bias but reduces variance

$A = \left(\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 \\ 1 \\ 0 \end{bmatrix} \right)$

The rank of $A = 2$ = the dim(span A) = # of linearly indep columns

$a1 + 2b = 5$

$a1 + 1b = 1$

$b = 4$

$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \\ 0 \end{bmatrix}$

- bonus info on subspaces S: Any set of vectors that satisfy the following

1.) $0 \in S$ 2.) $f, g \in S$ so $f + g \in S$ 3.) $f \in S$, $\alpha \in \mathbb{R}$, $\alpha f \in S$

- if your problem has no sol, we can approx it with a least squares sol

- Apply cross-validation to assess classifier performance (is this the mean)

Percent error = $\frac{\text{sum error vector}}{\text{length error vector}}$

when you keep a hold out set and use the rest of the data to train a model, then you test the accuracy of your classifier. (do this multiple times holding out a different portion of the data each time)

Then average the error rates