Lecture 6 EE 559 Jan 29, 2024

Machine Learning I: Supervised Methods

B. Keith Jenkins

Announcements

- Homework 2 is due Friday
- HW late submission policy posted
- Slido poll today
  - Trial period this week
  - Log on to slido.com
    - Provide your USC email address
    - Provide your name when asked
  - Slido poll number:
    - 3512316

D2L wk. 4 will move to "Syllabus and overall docs"

Reading

- Bishop 4.1.7 (perceptron)

Today's lecture

- Fundamental assumptions in supervised ML
- Vector and feature-space representations
  - Augmented notation, space
  - Distances in feature space
  - Weight space
  - Reflected data points → Lecture 7

Copyright 2021-2024 B. Keith Jenkins. All lecture notes contained herein are for use only by students and instructors of EE 559, Spring 2024.

EE 559 **Machine Learning I**
Jenkins

Fundamental Assumptions for ML (Classification and Regression)

1. **Sufficient information.** The information contained in the data, together with assumptions that are appropriate to the problem, are sufficient to permit generalization.
   - Less information in the data, requires more assumptions to be made.

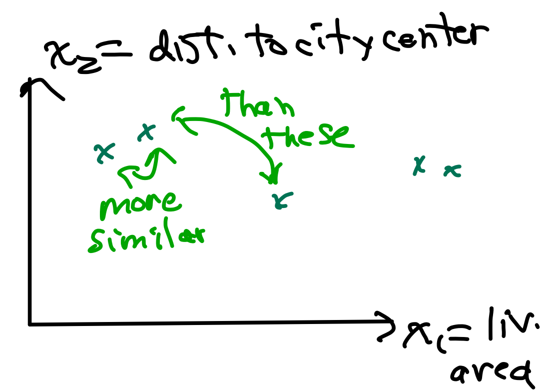
Assumptions → 1. Assumpt. on probability distributions of data.
2. " on features (e.g., what features would be relevant)

Tools → 3. Regularizers.
4. Priors. $P(S_1), P(S_2), \ldots$

2. **Feature-space representation.** There exists a representation and distance measure in feature space, such that the distance between data points represents their dis-similarity.
   - Thus, in feature space, closer points are more similar.

Can affect this:

- normalization
- distance measures
- transformation of feature space.
- kernels



3. **Input-to-output mapping.** There exists a correspondence between similarity of inputs, and similarity of outputs.

   – In classification, there is a similarity of instances from a given class, and a dis-similarity of instances from different classes. A class is a collection of instances with something in common.

     ✓ Ex1: predict students' grades in EE559

     ✗ Ex2: predict which random number is drawn by each student. (S_1: 0-25, S_2: 25-50, etc.)

   – In regression, there exists a function that can map from similarity of inputs to similarity of outputs.

4. **Representative data.** *Representative* sets of instances or data points are available (for training and testing). For supervised ML, these data points include output values.

   – The more similar the dataset is to the unknowns, the better.

   – If the dataset and unknowns are known to be dis-similar, other techniques (e.g., transfer learning) may be beneficial.

Augmented notation and space

→ For convenience

Discriminant fcns. $g(\underline{x})$, e.g. (linear case, 2-class): $g(\underline{x}) = \underline{w}^T \underline{x} + w_0$

→ **Def:**

$$\underline{w}^{(+)} = \underline{w}_+ = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}, \quad \underline{x}^{(+)} = \underline{x}_+ = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

$\underline{x}^{(+)} \Rightarrow$ augmented feature space.

Note: all data points $\underline{x}_n^{(+)} = \begin{bmatrix} \underline{1} \\ x_{n1} \\ x_{n2} \\ \vdots \\ x_{nD} \end{bmatrix}$

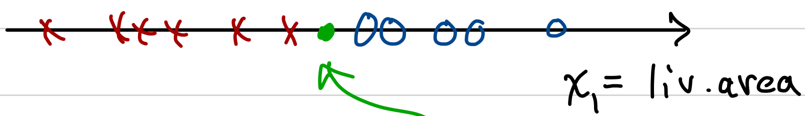
Then: $g(\underline{x}) = \underline{w}^{(+)T} \underline{x}^{(+)}$.

→ In slido question notation: $\begin{cases} \text{w\_0 means } w_0 \\ \text{w^T x means } w^T x \end{cases}$

Ex:

Let $D=1$.

Non-augm. feat space



decision boundary $x_1 = (x_1)_b$

x S_1: price increase

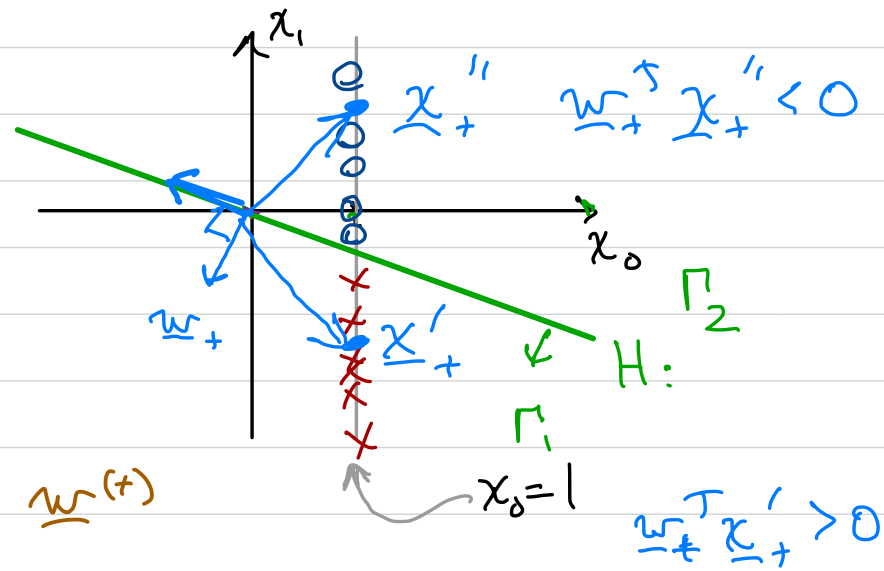
o S_2: price decrease

Eqn. for dec. boundary:

$$g(x) = 0 \Rightarrow w_1 x_1 + w_0 = 0$$

$$x_1 = \frac{-w_0}{w_1} = (x_1)_b$$

Augmented feat. space $\quad \underline{x} = \begin{bmatrix} 1 \\ x_1 \end{bmatrix}$



$$g(\underline{x}^{(+)}) = \underline{w}^{(+)T} \underline{x}^{(+)}$$

Dec. bound H:

$$g(\underline{0}) = 0$$

$\Rightarrow$ H must pass through origin.

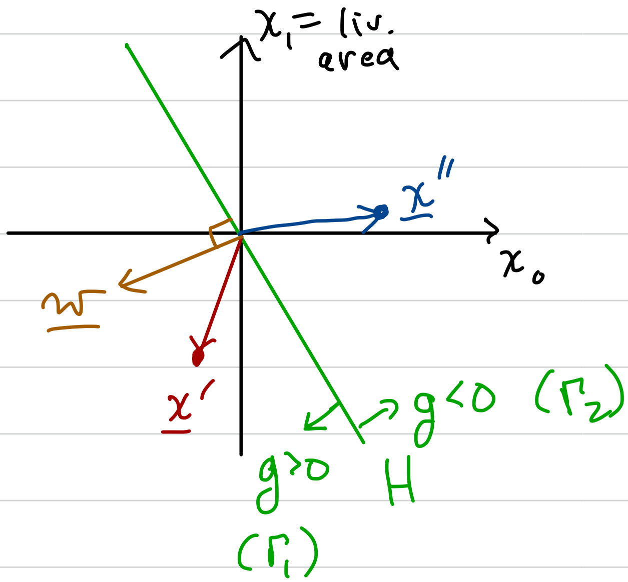
Also: for $\underline{x}^{(+)}$ on H:

$$g(\underline{x}^{(+)}) = \underline{w}^{(+)T} \underline{x}^{(+)} = 0$$

$$\Rightarrow \underline{w}^{(+)} \perp \underline{x}^{(+)}$$

Decision Regions

Direction of $\underline{w}$ in augm. feat. space (drop (+) superscript)

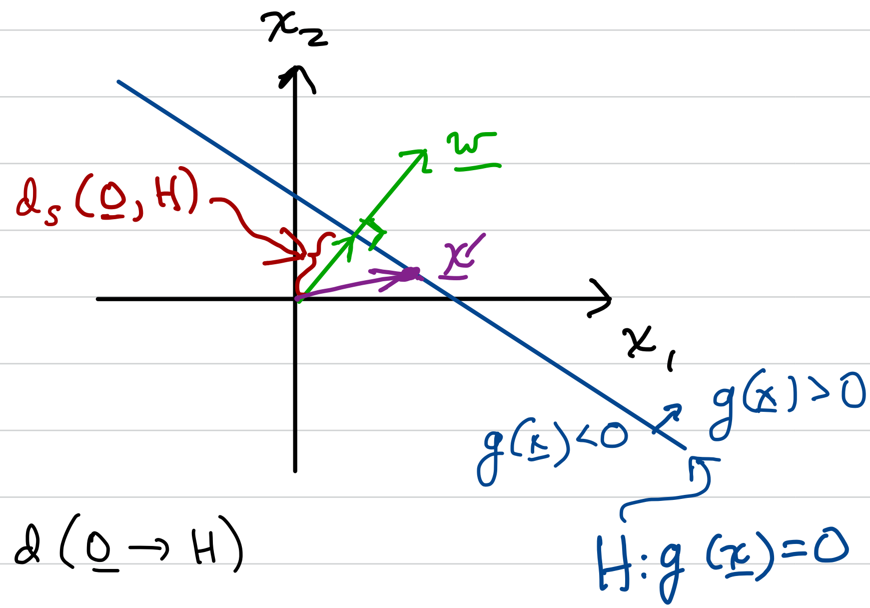


$$g(\underline{x}') = \underline{w}^T \underline{x}' > 0$$
$$g(\underline{x}'') = \underline{w}^T \underline{x}'' < 0$$

$\Rightarrow$ $\underline{w}$ points to $g>0$ (positive) side of H.

Distances in Feature Space (non-augmented)

(1) $g(\underline{x}) = w_0 + \underline{w}^T \underline{x}$



From multivariate calculus:
$\nabla_{\underline{x}} f(\underline{x}) \perp \{f(\underline{x}) = \text{const. curve}\}$ at any $\underline{x}$.

$\Rightarrow \nabla_{\underline{x}} g(\underline{x}) = \underline{w} \perp H: \; g(\underline{x}) = 0.$

"Signed distance" from $\underline{0}$ to H $= d_s(\underline{0}, H) = d(\underline{0} \to H)$

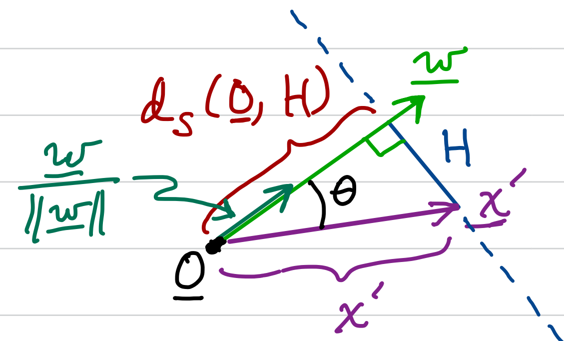
$$d_s(\underline{0}, H) = -d_s(H, \underline{0}).$$

Let $\underline{x}'$ be any point on H

Note: $\dfrac{\underline{w}}{\|\underline{w}\|}$ = unit vector in direction of $\underline{w}$.

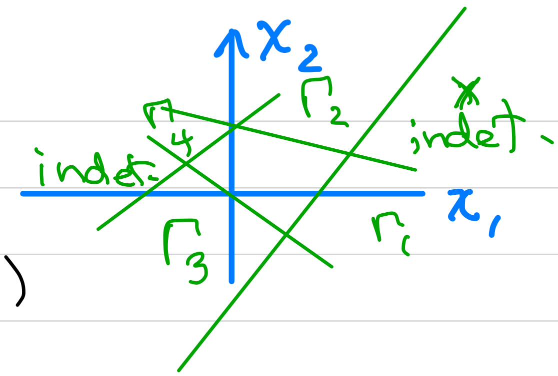
(2) $$d_s(\underline{0}, H) = \frac{\underline{w}^T}{\|\underline{w}\|}\underline{x}'$$

From: $\dfrac{\underline{w}^T}{\|\underline{w}\|}\underline{x}' = x' \cos\theta$
$= d_s(\underline{0}, H)$



$$= \frac{g(\underline{x}') - w_0}{\|\underline{w}\|} \quad \text{(from (1))}$$

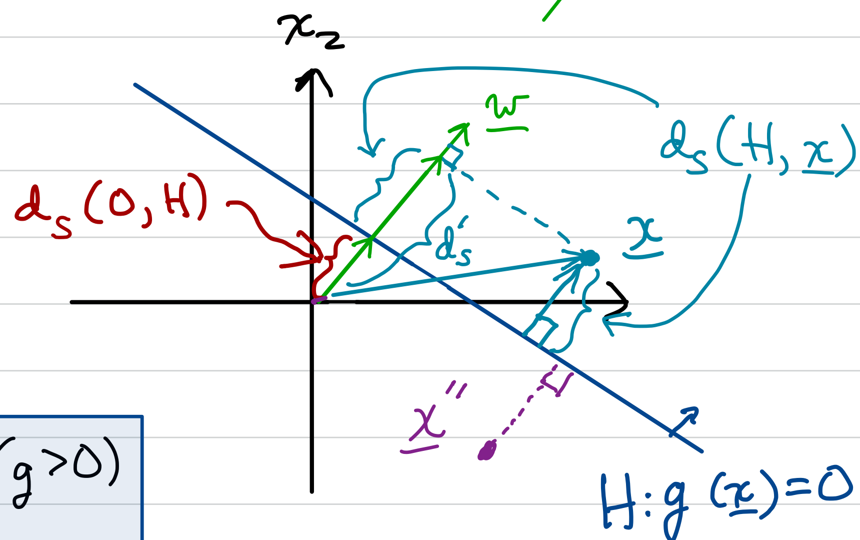
$$d_s(\underline{0}, H) = \frac{-w_0}{\|\underline{w}\|}$$



Find $d_s(H, \underline{x})$

$$d_s(H, \underline{x}) = \underbrace{d_s\{\underline{0} \text{ to } \underline{x} \text{ in direction} \perp H\}}_{d_s'} - d_s(\underline{0}, H)$$

$$= \frac{\underline{w}^T \underline{x}}{\|\underline{w}\|} - \frac{-w_0}{\|\underline{w}\|}$$



(3) $\begin{cases} \boxed{d_s(H, \underline{x}) = \dfrac{g(\underline{x})}{\|\underline{w}\|}} \\ \boxed{d_s(H, \underline{x}) > 0 \text{ if } \underline{x} \text{ is on positive } (g>0) \text{ side of } H.} \end{cases}$

(3) $\Rightarrow$ $\boxed{d_s(H, \underline{0}) = \dfrac{g(\underline{0})}{\|\underline{w}\|} = \dfrac{w_0}{\|\underline{w}\|}\ ; \quad d(H, \underline{0}) > 0 \text{ if } \underline{0} \text{ is on positive side of } H.}$

$$g(\underline{x}) = \underline{w}^T \underset{\uparrow\, \underline{0}}{\underline{x}} + w_0 = w_0$$

EE 559

Notation
(part 2)

Weight vector (non-augmented space):

$$\underline{w} = \left(w_1, \cdots, w_D\right)^T \quad \text{(general, or 2-class)}$$

$$\underline{w}_k = \underline{w}^{(k)} = \left(w_1^{(k)}, \cdots, w_D^{(k)}\right)^T \quad (\text{class } S_k,\ k = 1,2,\cdots,C;\ \ C > 2)$$

Linear discriminant function (2-class, non-augmented):

$$g\left(\underline{x}\right) = w_0 + \underline{w}^T \underline{x} = w_0 + \sum_{j=1}^{D} w_j x_j \qquad (\text{ Bishop } y\left(\underline{x}\right))$$

Linear discriminant functions (*C*-class, maximal value method, or OvR non-augmented):

$$g_k\left(\underline{x}\right) = w_0^{(k)} + \underline{w}_k^T \underline{x} = w_0^{(k)} + \sum_{j=1}^{D} w_j^{(k)} x_j \qquad (\text{ Bishop } y_k\left(\underline{x}\right))$$

Augmented and non-augmented feature space:

$$\underline{w}^{(0)} = \text{non-augmented weight vector} = \left(w_1, w_2, \cdots, w_D\right)^T \qquad (\text{Bishop } \underline{w})$$

$$\underline{x}^{(0)} = \text{non-augmented feature vector} = \left(x_1, x_2, \cdots, x_D\right)^T \qquad (\text{Bishop } \underline{x})$$

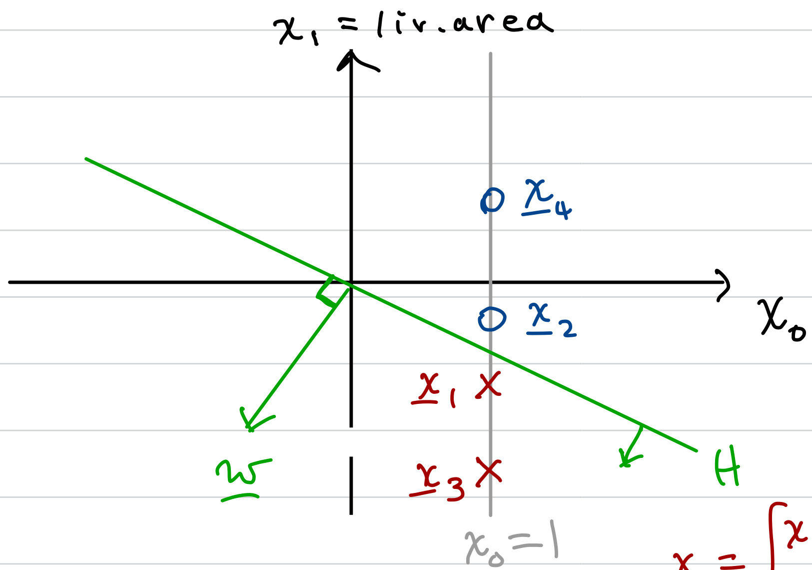
$$\underline{w}^{(+)} = \text{augmented weight vector} = \begin{bmatrix} w_0 \\ \underline{w}^{(0)} \end{bmatrix} \qquad (\text{Bishop } \tilde{\underline{w}})$$

$$\underline{x}^{(+)} = \text{augmented feature vector} = \begin{bmatrix} x_0 \\ \underline{x}^{(0)} \end{bmatrix} \begin{pmatrix} \text{any feature} \\ \text{space point} \end{pmatrix} \qquad (\text{Bishop } \tilde{\underline{x}})$$

$$\underline{x}_n^{(+)} = \text{augmented data vector} = \begin{bmatrix} 1 \\ \underline{x}^{(0)} \end{bmatrix} \begin{pmatrix} n^{\text{th}} \text{ data point} \\ \text{in feature space} \end{pmatrix} \qquad (\text{Bishop } \tilde{\underline{x}}_n)$$

Often we will omit the (+) and (0) superscripts, and instead state which space we are working in.

Feature space (augmented) D=1 feature



x S_1 price decr.
o S_2 price incr.

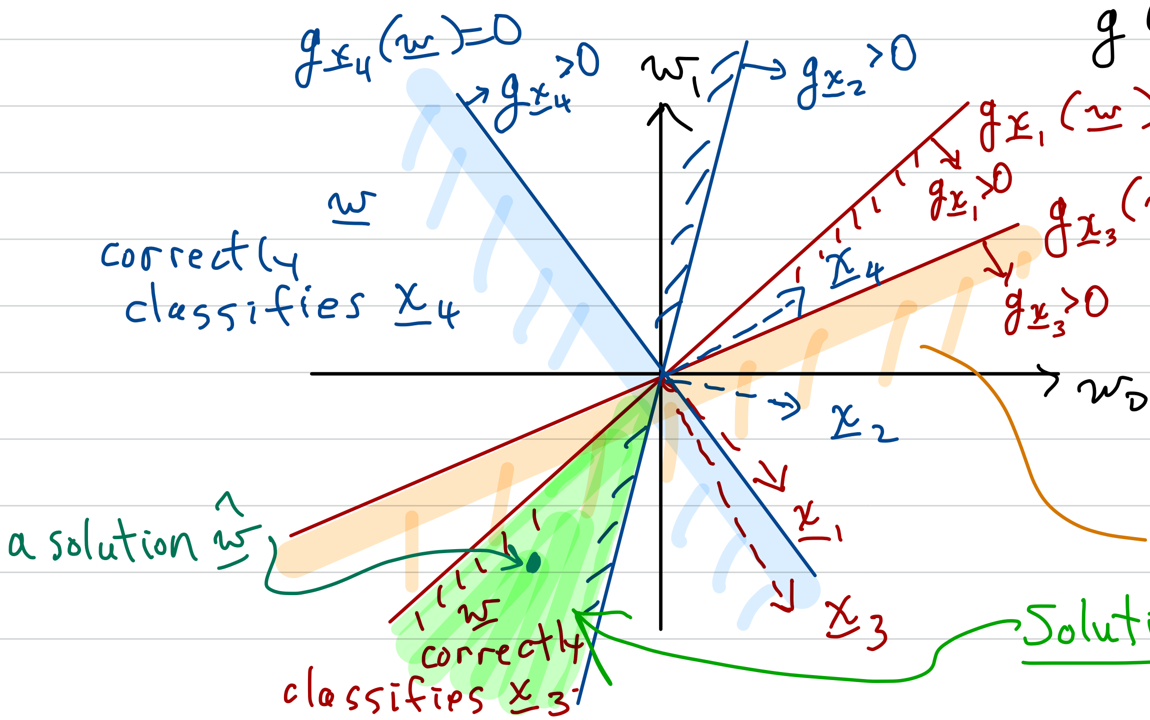
$$H:\ g(\underline{x}) = 0 = \underline{w}^T \underline{x}$$

$$H:\ w_0 x_0 + w_1 x_1 = 0$$

coefficients variables

$$\underline{x}_1 = \begin{bmatrix} x_{10} \\ x_{11} \end{bmatrix}$$

Weight space (augm.)



$$g(\underline{x}_1) = g_{\underline{x}_1}(\underline{w}) = x_{10} w_0 + x_{11} w_1$$

coefficients variables

$$g_{\underline{x}_1}(\underline{w}) = 0 \Rightarrow \text{line}$$

($\underline{w}$ for which $\underline{w}^T \underline{x}_1 = 0$)

$\underline{w}$ correctly classifies $\underline{x}_1$.

Solution region – all $\underline{w}$ correctly classify $\underline{x}_1, \underline{x}_2, \underline{x}_3, \underline{x}_4$.