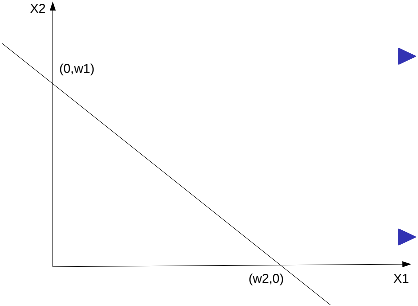
CS 474/574 Machine Learning
2. Linear Classifiers

Prof. Dr. Forrest Sheng Bao
Dept. of Computer Science
Iowa State University
Ames, IA, USA

September 4, 2020

The hyperplane



- Now, let's begin our journey on supervised learning.
- Suppose we have a line going thru points $(0, w_1)$ and $(w_2, 0)$ (which are the *intercepts*) in a 2-D vector space spanned by two orthogonal bases x_1 and x_2.
- The equation of this line is $x_1 w_1 + x_2 w_2 - w_1 w_2 = 0$.

The hyperplane (cond.)

- Let

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

  and

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ -w_1 w_2 \end{pmatrix}$$

  (By default, all vectors are column vectors.)
- x_1 and x_2 are two **feature values** comprising the feature vector. 1 is **augmented** for the bias $-w_1 w_2$. Then the equation is rewritten into vector form: $\mathbf{x}^T \cdot \mathbf{w} = 0$. For space sake, $\mathbf{x}^T \mathbf{w} = \mathbf{x}^T \cdot \mathbf{w}$.

The hyperplane (cond.)

- Expand to n-dimension.

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{pmatrix}$$

and

$$\mathbf{W} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ -w_1 w_2 \end{pmatrix}.$$

Then $\mathbf{X}^T \cdot \mathbf{W} = 0$, denoted as the *hyperplane* in $\mathbb{R}^n$.

Binary Linear Classifier

- A binary linear classier is a function $f(X) = \mathbf{W}\mathbf{X}$, such that

$$\begin{cases} \mathbf{W}^T\mathbf{X} > 0 & \forall X \in C_1 \\ \mathbf{W}^T\mathbf{X} < 0 & \forall X \in C_2 \end{cases} \tag{1}$$

  where C_1 and C_2 are the two classes. Note that the $\mathbf{X}$ has been augmented with 1 as mentioned before.
- Using the function f to make decision is called *test*. Given a new sample whose augmented feature vector is $\mathbf{X}$, if $\mathbf{W}^T\mathbf{X} > 0$, then we classify the sample to class C_1. Otherwise, class C_2.
- Example. Let $\mathbf{W}^T = (2, 4, -8)$, what's the class for new sample $\mathbf{X} = (1, 1, 1)$ (1 is augmented)?
- $\mathbf{W}^T\mathbf{X} = -2 < 0$. Hence the sample of feature value $(1, 1)$ belongs to class C_1.

Normalized feature vector

- Eq. 1 has two directions. Let's unify them into one.
- A correctly classified sample $(\mathbf{X_i}, y_i)$ shall satisfy the inequality $\mathbf{W}_i^T \mathbf{X} y_i > 0$. (The y_i flips the direction of the inequality.)
- *normalize* the feature vector: $\mathbf{X}_i y_i$ for $y_i \in \{+1, -1\}$.
- Example:
  - $\mathbf{x}'_1 = (0, 0)^T$, $\mathbf{x}'_2 = (0, 1)^T$, $\mathbf{x}'_3 = (1, 0)^T$, $\mathbf{x}'_4 = (1, 1)^T$,
  - $y_1 = 1, y_2 = 1, y_3 = -1, y_4 = -1$
  - First, augment them: $\mathbf{x}_1 = (0, 0, 1)^T$, $\mathbf{x}_2 = (0, 1, 1)^T$, $\mathbf{x}_3 = (1, 0, 1)^T$, $\mathbf{x}_4 = (1, 1, 1)^T$
  - Then, normalize them $\mathbf{x}'_1 = \mathbf{x}_1$, $\mathbf{x}'_2 = \mathbf{x}_2$, $\mathbf{x}'_3 = -\mathbf{x}_3 = (-1, 0, -1)^T$, $\mathbf{x}'_4 = \mathbf{x}_4 = (-1, -1, -1)^T$
- Please note that the term ''normalized" could have different meanings in different context of ML.

Solving inequalities: the simplest way to find the $\mathbf{W}$

- Let's look at a case where the feature vector is 1-D.
- Let the training set be $\{(4, C_1), (5, C_1), (1, C_2), (2, C_2)\}$. Their augmented feature vectors are: $X_1 = (4, 1)^T$, $X_2 = (5, 1)^T$, $X_3 = (1, 1)^T$, $X_4 = (2, 1)^T$.
- Let $\mathbf{W}^T = (w_1, w_2)$. In the training process, we can establish 4 inequalities:

$$\begin{cases} 4w_1 + w_2 & > 0 \\ 5w_1 + w_2 & > 0 \\ w_1 + w_2 & < 0 \\ 2w_1 + w_2 & < 0 \end{cases}$$

- We can find many w_1 and w_2 to satisfy the inequalities. But, how to pick the best?

Math recap: Gradient

- The partial derivative of a multivariate function is a vector called the gradient, representing the derivatives of a function on different directions.
- For example, let $f(\mathbf{x}) = x_1^2 + 4x_1 + 2x_1x_2 + 2x_2^2 + 2x_2 + 14$. f maps a vector $\mathbf{x} = (x_1, x_2)^T$ to a scalar.
- Then we have

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2x_1 + 2x_2 - 4 \\ 4x_2 + 2x_1 + 2 \end{pmatrix}$$

- The gradient is a special case of *Jacobian matrix*. (see also: *Hessian matrix* for second-order partial derivatives.)
- A *critical point* or a *stationary point* is reached where the derivative is zero on any direction.
  - a local extremum (a maximum or a minimum)
  - saddle point
- if a function is convex, a local minimum/maxinum is the *global minimum/maximum*.

Finding the linear classifier via zero-gradient

- Two steps here:
  - Define a cost function to be minimized (The learning is the about minimizing the cost function)
  - Choose an algorithm to minimize (e.g., gradient, least squared error etc.)
- One intuitive criterion can be the sum of error square:

$$J(\mathbf{W}) = \sum_{i=1}^{N} (\mathbf{W}^T \mathbf{x}_i - y_i)^2 = \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{W} - y_i)^2$$

Finding the linear classifier via zero-gradient (cond.)

- Minimizing $J(\mathbf{W})$ means (Convexity next time.)
  $\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}} = 2\sum_{i=1}^{N} \mathbf{x}_i(\mathbf{x}_i^T\mathbf{W} - y_i) = (0, \ldots, 0)^T$
- Hence, $\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T\mathbf{W} = \sum_{i=1}^{N} \mathbf{x}_i y_i$
- The sum of a column vector multiplied with a row vector produces a matrix.

$$\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T = \begin{pmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ | & | & & | \end{pmatrix} \begin{pmatrix} — & \mathbf{x}_1^T & — \\ — & \mathbf{x}_2^T & — \\ & \vdots & \\ — & \mathbf{x}_N^T & — \end{pmatrix} = \mathbb{X}^T\mathbb{X}$$

Finding the linear classifier via zero-gradient (cond.)

- $$\sum_{i=1}^{N} \mathbf{x}_i y_i = \begin{pmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ | & | & & | \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \mathbb{X}^T \mathbf{y}$$
- $\mathbb{X}^T \mathbb{X} \mathbf{W} = \mathbb{X}^T \mathbf{y}$
- $(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{X} \mathbf{W} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{y}$
- $\mathbf{W} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{y}$

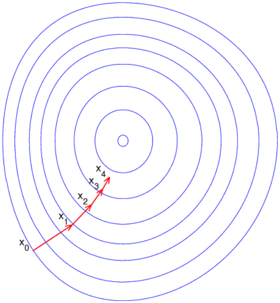
Gradient descent approach

Since we define the target function as $J(\mathbf{W})$, finding $J(\mathbf{W}) = 0$ or minimizing $J(\mathbf{W})$ is intuitively the same as reducing $J(\mathbf{W})$ along the gradient. The algorithm below is a general approach to minimize any multivariate function: changing the input variable proportionally to the gradient.

Algorithm 1: pseudocode for gradient descent approach

Input: an initial $\mathbf{w}$, stop criterion θ, a learning rate function $\rho(\cdot)$, iteration step $k = 0$

while $\nabla J(\mathbf{w}) > \theta$ **do**
    $\mathbf{w}_{k+1} := \mathbf{w}_k - \rho(k)\nabla J(\mathbf{w})$
    $k := k + 1$
end while



Gradient descent approach (cond.)

In many cases, the $\rho(k)$'s amplitude (why amplitude but not the value?) decreases as k increases, e.g., $\rho(k) = \frac{1}{k}$, in order to shrink the adjustment.Also in some cases, the stop condition is $\rho(k)\nabla J(\mathbf{w}) > \theta$. The limit on k can also be included in stop condition – do not run forever.

Fisher's linear discriminant

- What really is $\mathbf{w}^T x$? $\mathbf{w}$ is perpendicular to the hyper panel [^3]
- $\mathbf{w}^T \mathbf{x}$ is the *projection* of the point $\mathbf{x}$ on the decision panel.
- Intuition in a simple example: for any two points $\mathbf{x}_1 \in C_1$ and $\mathbf{x}_2 \in C_2$, we want $\mathbf{w}^T \mathbf{x}_1$ to be as different from $\mathbf{w}^T \mathbf{x}_1$ as possible, i.e., $\max(\mathbf{w}^T \mathbf{x}_1 - \mathbf{w}^T \mathbf{x}_2)^2$. [Fig. 4.6, Bishop book]
- For binary classification, intuitively, we want the projections of the same class to be close to each other (i.e., the smaller $\tilde{s}_1$ and $\tilde{s}_2$ the better) while the projects of different classes to be apart from each other (i.e., the larger $(\tilde{m}_1 - \tilde{m}_2)^2$ is better).
- That means

$$\max J(\mathbf{w}) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

where $\tilde{m}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{w}^T \mathbf{x}$ and $\tilde{\mathbf{s}}_i^2 = \sum_{\mathbf{x} \in C_i} (\mathbf{w}^T \mathbf{x} - \tilde{m}_i)^2$ are the mean and the variance of the projection of all samples belonging to Class i on the decision panel, respectively.

Fisher's (cond.)

- between-class scatter:
  $(\tilde{m}_1 - \tilde{m}_2)^2 = (\mathbf{w}^T(\mathbf{m_1} - \mathbf{m_2}))^2 = \mathbf{w}^T(\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^T\mathbf{w}$
  where $\mathbf{m}_i = \frac{1}{|C_i|}\sum_{\mathbf{x}\in C_i}\mathbf{x}$
- within-class scatter: $\tilde{\mathbf{s}}_i^2 = \sum_{\mathbf{x}\in C_i}(\mathbf{w}^T\mathbf{x} - \tilde{m}_i)^2 =$
  $\sum_{\mathbf{x}\in C_i}(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i)^2 = \mathbf{w}^T[\sum_{\mathbf{x}\in C_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i^T)]\mathbf{w}$
- Denote $\mathbf{S_w} = \tilde{\mathbf{s}}_1^2 + \tilde{\mathbf{s}}_2^2$ and $\mathbf{S}_B = (\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^T$. We have

  $$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$

  . This expression is known as *Rayleigh quotient*.
- To maximize $J(\mathbf{w})$, the $\mathbf{w}$ must satisfy $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$.
- Hence $\mathbf{w} = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$. (Derivate saved.)