Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 3 - Linear Regression - Probabilistic Interpretation and Regularization

Recap: Linear Regression is **not Naively Linear**

Linear in w_i's

- Need to determine $\mathbf{w}$ for the linear function $f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} w_i \phi_i(\mathbf{x_j}) = \cancel{\Phi \mathbf{w}}$ which minimizes our error function $E\left(f(\mathbf{x}, \mathbf{w}), \mathcal{D}\right)$

  $\phi^T w$

Extremely non-linear in x's.

- Owing to basis function ϕ, "Linear Regression" is *linear* in $\mathbf{w}$ but NOT in $\mathbf{x}$ (which could be arbitrarily non-linear)!

$$\Phi = \begin{bmatrix} \phi_1(\mathbf{x_1}) & \phi_2(\mathbf{x_1}) & \ldots\ldots & \phi_n(\mathbf{x_1}) \\ \cdot & & & \\ \cdot & & & \\ \phi_1(\mathbf{x_m}) & \phi_2(\mathbf{x_m}) & \ldots\ldots & \phi_n(\mathbf{x_m}) \end{bmatrix} \quad (1)$$

Recap: Linear Regression is **not Naively Linear**

- Need to determine $\mathbf{w}$ for the linear function $f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} w_i \phi_i(\mathbf{x}_j) = \cancel{\Phi}^T\mathbf{w}$ $\phi^T w$ which minimizes our error function $E(f(\mathbf{x}, \mathbf{w}), \mathcal{D})$
- Least Squares error and corresponding estimates:

Sum across rows of Φ — sum within row of Φ

$$E^* = \min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \underbrace{\left\{ \sum_{j=1}^{m} \left(\sum_{i=1}^{n} w_i \phi_i(\mathbf{x}_j) - y_j \right)^2 \right\}}_{\text{in matrix form?}} \quad (2)$$

E^* is attained at

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \arg\min_{w} (\Phi w - y)^T (\Phi w - y)$$

On test point x, use $f(x, w^*) = $ fn of $\mathcal{D}$ & choice of ϕ ...

Recap: Linear Regression is **not Naively Linear**

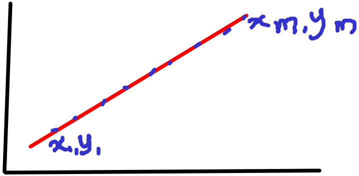
- Need to determine $\mathbf{w}$ for the linear function $f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} w_i \phi_i(\mathbf{x}_j) = \Phi\mathbf{w}$ which minimizes our error function $E(f(\mathbf{x}, \mathbf{w}), \mathcal{D})$
- Least Squares error and corresponding estimates:

$$E^* = \min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \left\{ \sum_{j=1}^{m} \left(\sum_{i=1}^{n} w_i \phi_i(\mathbf{x}_j) - y_j \right)^2 \right\} \tag{2}$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \arg\min_{\mathbf{w}} \underbrace{\left(\mathbf{w}^T \Phi^T \Phi \mathbf{w} - 2\mathbf{y}^T \Phi \mathbf{w} + \mathbf{y}^T \mathbf{y} \right)}_{\|\Phi w - y\|_2^2} \tag{3}$$

- The minimum value of the squared loss is zero
- If zero were attained at $\mathbf{w}^*$, we would have $\forall j, \phi^T(x_j)\mathbf{w}^* = y_j$, or equivalently $\Phi\mathbf{w}^* = \mathbf{y}$, where

$$E = ||\Phi w^* - y||^2 = 0$$



$$\Phi = \begin{array}{c} \phi^T(x_1) \\ \begin{bmatrix} \phi_1(x_1) & \dots & \phi_p(x_1) \\ \dots & \dots & \dots \\ \phi_1(x_m) & \dots & \phi_p(x_m) \end{bmatrix} \\ \phi^T(x_m) \end{array}$$

and

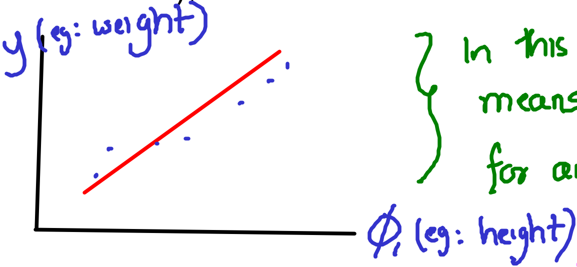
$$\mathbf{y} = \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix}$$

What if there is no perfect line fit?

- It has a solution if $\mathbf{y}$ is in the column space (the subspace of $\Re^m$ formed by the column vectors) of Φ

- The minimum value of the squared loss is zero
- If $\mathbf{y}$ is NOT in the column space of Φ (that is, if zero is NOT attainable at $\mathbf{w}^*$), what can be done?

$$\Phi = \begin{bmatrix}\phi_1(x_1)\\ \phi_1(x_2)\\ \phi_1(x_m)\end{bmatrix}$$



In this 1-d case, this means $y \neq \alpha \Phi$ for any $\alpha \in \mathbb{R}$

Soln: Let $y^* = \alpha^* \Phi$ s.t y^* is closest to y

Geometric Interpretation of Least Square Solution

- Let $\mathbf{y}^*$ be a solution in the column space of Φ
- The least squares solution is such that the distance between $\mathbf{y}^*$ and $\mathbf{y}$ is minimized
- Therefore...........

If # of pts increases beyond 2, the number of coordinates in the figure will increase

If # of features (ϕ_i's) increases beyond 1, the dimensionality of $c(\phi)$ will grow to a plane

For 1-dim case, $\Phi = \begin{bmatrix}\phi_1(x_1)\\ \vdots \\ \phi_1(x_m)\end{bmatrix} \in \mathbb{R}^m$

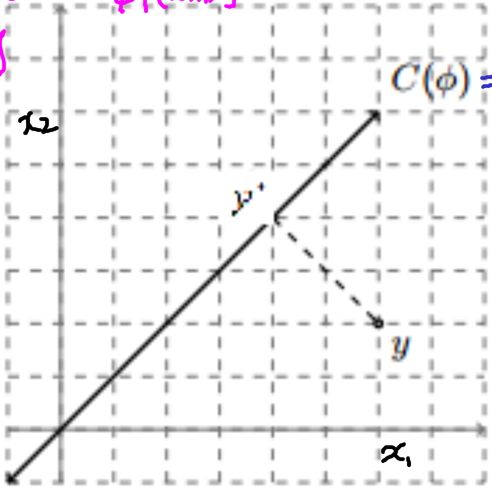
$C(\Phi) = \{\Phi\alpha \mid \alpha \in \mathbb{R}\}$

= one dim vector in $\mathbb{R}^m$

$y^* \in \mathbb{R}^m$

$y \in \mathbb{R}^m$

$C(\phi) = \{\Phi w \mid w \in \mathbb{R}^n\}$



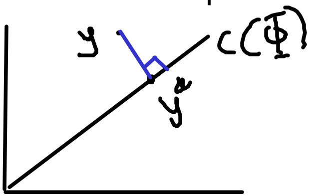
y will be closest to y^* when $(y - y^*) \perp C(\phi)$

This plot depicts $C(\cdot)$ for 1 feature ϕ_1 and 2 data points: x_1 & x_2 $(m=2)$

Geometric Interpretation of Least Square Solution

4
- Let $\mathbf{y}^*$ be a solution in the column space of Φ $[y^* = \Phi w^* \in C(\Phi)]$
- The least squares solution $\mathbf{w}^*$ is such that the distance between $\mathbf{y}^*$ and $\mathbf{y}$ is minimized

5
- Therefore, the line joining $\mathbf{y}^*$ to $\mathbf{y}$ should be orthogonal to the column space



$$\Phi \mathbf{w}^* = \mathbf{y}^* \tag{4}$$

$$(\mathbf{y} - \mathbf{y}^*)^T \Phi = 0 \tag{5}$$

$(y-y^*) \perp C(\Phi) = (y-y^*) \perp \text{span}(\text{columns of } \Phi)$

$= (y-y^*) \perp \Phi$

$$(\mathbf{y}^*)^T \Phi = (\mathbf{y})^T \Phi \tag{6}$$

$(y^{\wedge})^T \Phi = y^T \Phi$ $\longrightarrow$

From (4), $y^{\wedge} = \Phi w$

Substitute using (4)

$$(\Phi \mathbf{w})^T \Phi = \mathbf{y}^T \Phi \tag{7}$$

$(AV)^T = V^T A^T$

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

$$\mathbf{w}^T \Phi^T \Phi = \mathbf{y}^T \Phi \tag{8}$$

Taking transposes on both sides

$$\Phi^T \Phi \mathbf{w} = \Phi^T \mathbf{y} \tag{9}$$

$$\mathbf{w} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} \tag{10}$$

- Here $\Phi^T \Phi$ is invertible if and only if Φ has full column rank

ie. no feature set for one data point can be obtained using the feature sets of other data points

$$(\Phi \mathbf{w})^T \Phi = \mathbf{y}^T \Phi \tag{7}$$

$$\mathbf{w}^T \Phi^T \Phi = \mathbf{y}^T \Phi \tag{8}$$

$$\Phi^T \Phi \mathbf{w} = \Phi^T \mathbf{y} \tag{9}$$

$$\mathbf{w} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} \tag{10}$$

- Here $\Phi^T \Phi$ is invertible if and only if Φ has full column rank

PROOF?

Proof:

Claim : $\Phi^T\Phi$ is invertible if and only if Φ is full column rank

Proof :

$\Longleftarrow$ Given that Φ has full column rank and hence columns are linearly independent, we have that $\Phi\mathbf{w} = 0 \Rightarrow \mathbf{w} = 0$ (only vector in null space is 0)

Assume on the contrary that $\Phi^T\Phi$ is non invertible. Then $\exists\mathbf{w} \neq 0$ such that $\Phi^T\Phi\mathbf{w} = 0$

$$\Rightarrow \mathbf{w}^T\Phi^T\Phi\mathbf{w} = 0$$

(just premultiplying 0 vec by 0)

$(\phi w)^T$

$$\Rightarrow (\Phi\mathbf{w})^T\Phi\mathbf{w} = 0$$

$$\Rightarrow \Phi\mathbf{w} = 0$$

$\|\phi w\|^2 = 0$ iff $\phi w = 0$

This is a contradiction. Hence $\Phi^T\Phi$ is invertible if Φ is full column rank

Claim : $\Phi^T\Phi$ is invertible if and only if Φ is full column rank

Proof :

$\Longrightarrow$ If $\Phi^T\Phi$ is invertible and $\Phi\mathbf{w} = 0$, then $(\Phi^T\Phi\mathbf{w}) = 0$, which in turn implies $\mathbf{w} = 0$. This implies Φ has full column rank if $\Phi^T\Phi$ is invertible. The converse can also be proved similarly.

Later: More on Optimization

$w^* = (\phi^T\phi)^{-1}\Phi^T y$ minimizes $E(\cdot) = \|\phi w - y\|_2^2$

- More generally: How to minimize a function?
    - Level Curves and Surfaces
    - Gradient Vector
    - Directional Derivative
    - Hyperplane
    - Tangential Hyperplane
- Iterative Algorithms such as (Stochastic) Gradient Descent Algorithm

Building on questions on Least Squares Linear Regression

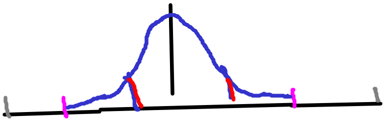
1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates, Regularization
3. How to minimize the resultant and more complex error functions?
   - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Building on questions on Least Squares Linear Regression

1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates, Regularization
3. How to minimize the resultant and more complex error functions?
   - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Probabilistic Modeling of Linear Regression

- Linear Model: Y is a linear function of $\phi(x)$, subject to a random noise variable ε which we believe is 'mostly' bounded by some threshold σ:



$$Y = w^T\phi(x) + \varepsilon$$

Random noise

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$H(p) = -\int_x p(x) \ln p(x)\, dx = E[-\ln p(x)]$$

- Motivation: $\mathcal{N}(\mu, \sigma^2)$, has maximum entropy among all real-valued distributions with a specified variance σ^2
- $3-\sigma$ rule: About 68% of values drawn from $\mathcal{N}(\mu, \sigma^2)$ are within one standard deviation σ away from the mean μ; about 95% of the values lie within 2σ; and about 99.7% are within 3σ.

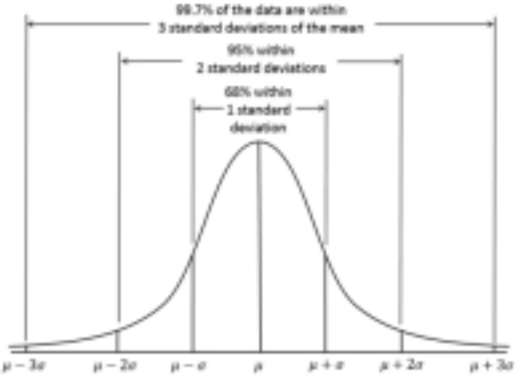


Figure 2: $3-\sigma$ rule: About 68% of values drawn from $\mathcal{N}(\mu, \sigma^2)$ are within one standard deviation σ away from the mean μ; about 95% of the values lie within 2σ; and about 99.7% are within 3σ. Source: `https://en.wikipedia.org/wiki/Normal_distribution`

Probabilistic Modeling of Linear Regression

- Linear Model: Y is a linear function of $\phi(\mathbf{x})$, subject to a random noise variable ε which we believe is 'mostly' around some threshold σ:

shifted Gaussian

$$Y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

shifted mean

- This allows for the Probabilistic model

$$P(y_j | \mathbf{w}, \mathbf{x}_j, \sigma^2) = \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}_j), \sigma^2)$$

$$P(y | \mathbf{w}, \mathbf{x}_j, \sigma^2) = \prod_{j=1}^{m} P(y_j | \mathbf{w}, \mathbf{x}_j, \sigma^2)$$

- Note: $E[Y(\mathbf{w}, \mathbf{x}_j)] =$ $w^T \phi(x)$

Probabilistic Modeling of Linear Regression

- Linear Model: Y is a linear function of $\phi(\mathbf{x})$, subject to a random noise variable ε which we believe is 'mostly' around some threshold σ:

$$Y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- This allows for the Probabilistic model

$$P(y_j | \mathbf{w}, \mathbf{x}_j, \sigma^2) = \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}_j), \sigma^2)$$

$$P(y | \mathbf{w}, \mathbf{x}_j, \sigma^2) = \prod_{j=1}^{m} P(y_j | \mathbf{w}, \mathbf{x}_j, \sigma^2)$$

- Note: $E[Y(\mathbf{w}, \mathbf{x}_j)] = \mathbf{w}^T \phi(\mathbf{x}_j)$
$= \mathbf{w}_0^T + \mathbf{w}_1^T \phi_1(\mathbf{x}_j) + \ldots + \mathbf{w}_n^T \phi_n(\mathbf{x}_j)$