Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 4 - Linear Regression - Probabilistic Interpretation and Regularization

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

- Need to determine $\mathbf{w}$ for the linear function $f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} w_i \phi_i(\mathbf{x_j}) = \mathbf{\Phi w}$ which minimizes our error function $E(f(\mathbf{x}, \mathbf{w}), \mathcal{D})$
- Owing to basis function ϕ, "Linear Regression" is *linear* in $\mathbf{w}$ but NOT in $\mathbf{x}$ (which could be arbitrarily non-linear)!

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

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

- Need to determine **w** for the linear function $f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} w_i \phi_i(\mathbf{x_j}) = \mathbf{\Phi w}$ which minimizes our error function $E(f(\mathbf{x}, \mathbf{w}), \mathcal{D})$
- Owing to basis function ϕ, "Linear Regression" is *linear* in **w** but NOT in **x** (which could be arbitrarily non-linear)!

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

- Least Squares error and corresponding estimates:

$$E^* = \min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \left(\mathbf{w^T \Phi^T \Phi w - 2y^T \Phi w + y^T y} \right) \quad (2)$$

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

Recap: 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, the line joining $\mathbf{y}^*$ to $\mathbf{y}$ should be orthogonal to the column space of $\Phi \Rightarrow$

$$\mathbf{w} = (\boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathbf{T}}\mathbf{y} \tag{4}$$

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

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:

$$Y = w^T \phi(x) + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- 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.

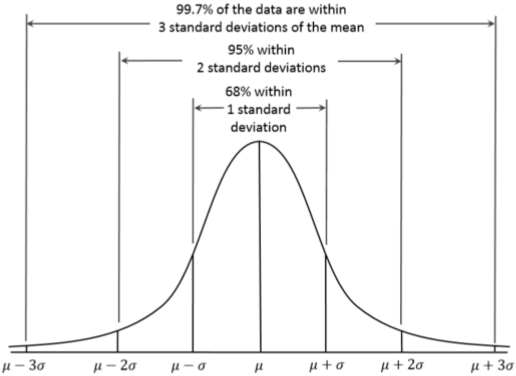

Figure 1: $3 - \sigma$ rule: About 68% of values drawn from $\mathcal{N}(\mu, \sigma^2)$ are within one standard deviation σ away from the mean μ; about 95% of the values lie within 2σ; and about 99.7% are within 3σ. Source: `https://en.wikipedia.org/wiki/Normal_distribution`

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:

$$Y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- This allows for the Probabilistic model

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

- Another motivation: $E[Y(\mathbf{w}, \mathbf{x}_j)] =$

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

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

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

Estimating $\mathbf{w}$: Maximum Likelihood

- If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y = \mathbf{w}^\mathbf{T}\phi(\mathbf{x}) + \epsilon$ where $\mathbf{w},\ \phi(\mathbf{x}) \in \mathbf{R}^\mathbf{m}$ then, given dataset $\mathcal{D}$, find the most likely $\hat{\mathbf{w}_{ML}}$
- Recall: $\Pr(y_j|\mathbf{x}_j, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(y_j - \mathbf{w}^T\phi(\mathbf{x}_j))^2}{2\sigma^2}\right)$
- From *Probability of data* to *Likelihood of parameters*: $\Pr(\mathcal{D}|\mathbf{w}) = \Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) =$

Estimating $\mathbf{w}$: Maximum Likelihood

- If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y = \mathbf{w}^\mathbf{T}\phi(\mathbf{x}) + \epsilon$ where $\mathbf{w},\ \phi(\mathbf{x}) \in \mathbf{R}^\mathbf{m}$ then, given dataset $\mathcal{D}$, find the most likely $\hat{\mathbf{w}_{ML}}$
- Recall: $\Pr(y_j|\mathbf{x}_j, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(y_j - \mathbf{w}^T\phi(\mathbf{x}_j))^2}{2\sigma^2}\right)$
- From *Probability of data* to *Likelihood of parameters*: $\Pr(\mathcal{D}|\mathbf{w}) = \Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) =$ $\prod_{j=1}^{m} \Pr(y_j|\mathbf{x}_j, \mathbf{w}) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(y_j - \mathbf{w}^T\phi(\mathbf{x}_j))^2}{2\sigma^2}\right)$
- Maximum Likelihood Estimate $\hat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}} \Pr(\mathcal{D}|\mathbf{w}) = \Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) = L(\mathbf{w}|\mathcal{D})$

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as $\log$)

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as log)
- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$
  $-\frac{m}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w^T}\phi(\mathbf{x_j}) - \mathbf{y_j})^\mathbf{2}$
  For a fixed σ^2
  $\mathbf{w}\hat{_{ML}} =$

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as log)
- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$
  $$-\frac{m}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w^T}\phi(\mathbf{x_j}) - \mathbf{y_j})^\mathbf{2}$$
  For a fixed σ^2
  $$\hat{\mathbf{w}_{ML}} = \text{argmax } LL(y_1 ... y_m|\mathbf{x}_1 \dots \mathbf{x}_m, \mathbf{w}, \sigma^2)$$
  $$= \text{argmin } \sum_{j=1}^{m}(\mathbf{w}^T\phi(\mathbf{x}_j) - y_j)^2$$
- Note that this is same as the Least square solution!!

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

Redundant Φ and Overfitting

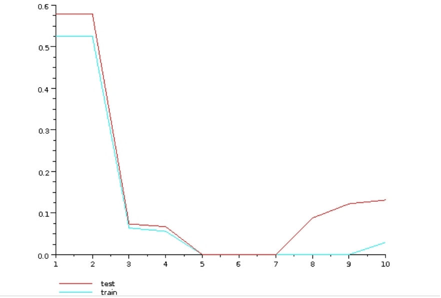

Figure 2: Root Mean Squared (RMS) errors on sample train and test datasets as a function of the degree t of the polynomial being fit

- Too many bends (t=9 onwards) in curve $\equiv$ high values of some $w_i^{\prime}s$. Try plotting values of w_i's using applet at

  http://mste.illinois.edu/users/exner/java.f/leastsquares/#simulation

- Train and test errors differ significantly

Bayesian Linear Regression

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief**
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression
- Continue with Normally distributed errors
- Model the $\mathbf{w}$ using a prior distribution and use the posterior over $\mathbf{w}$ as the result
- Intuitive Prior:

Bayesian Linear Regression

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief**
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression
- Continue with Normally distributed errors
- Model the $\mathbf{w}$ using a prior distribution and use the posterior over $\mathbf{w}$ as the result
- Intuitive Prior: Components of $\mathbf{w}$ should not become too large!
- Next: Illustration of Bayesian Estimation on a simple Coin-tossing example