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)!

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

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

function $f(x,w)$ in x

- 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)!

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

- Least Squares error and corresponding estimates:

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

Derived graphically

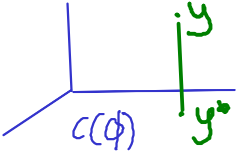
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbf{E}(\mathbf{w},\mathcal{D}) = \arg\min_{\mathbf{w}} \left\{\sum_{\mathbf{j=1}}^{\mathbf{m}}\left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i}\phi_{\mathbf{i}}(\mathbf{x_j}) - \mathbf{y_j}\right)^2\right\} \tag{3}$$

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 Φ
  $\Rightarrow$

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

- 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:

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

(exponential, χ^2, t, uniform)

- 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.

p^* is a distr to be estimated

$\mathcal{P}$ is family of pdfs $\supseteq \{ N, \text{exponential}, t, \chi^2, \ldots \}$

$p^* = $ maximize scope $\equiv \max_{p} \underbrace{\int_x \left(-\log_2 p(x)\right) p(x)\, dx}_{\text{Budgeted encoding using } p}$

s.t $\text{var}_{p^*}[X] = \sigma^2$

more prob $\Rightarrow$ less specified

distr over outcomes $\{$day, hour, mins, secs $\ldots\}$

more prob $\Rightarrow$ more imp

increasing imp

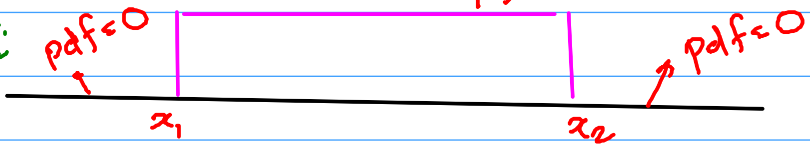
Q: What will p^* be if $\mathcal{P}$ is family of discrete distributions?

$$\underset{p_1, p_2 \ldots p_k}{\arg\max} \sum_i -p_i \log_2 p_i = \left\{ p_1^* = p_2^* \cdots = p_k^* = \frac{1}{k} \right\}$$

Why isn't uniform the entropy maximizer in the continuous case?

pdf $\dfrac{1}{x_2 - x_1}$

Ans:



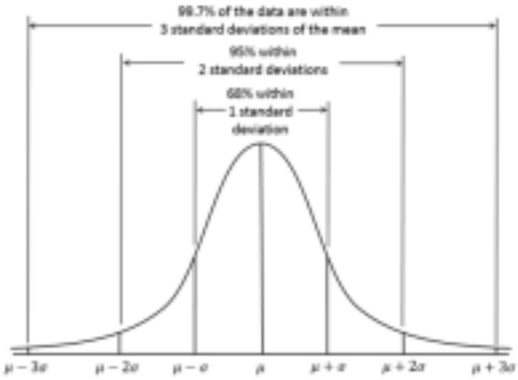


Figure 1: $3 - \sigma$ rule: About 68% of values drawn from $\mathcal{N}(\mu, \sigma^2)$ are within one standard deviation σ away from the mean μ; about 95% of the values lie within 2σ; and about 99.7% are within 3σ. Source: `https://en.wikipedia.org/wiki/Normal_distribution`

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

$\underbrace{Y = f + \varepsilon}$

$E\left[e^{f+\varepsilon}\right]_{\varepsilon \sim N(0,\sigma^2)}$

$E(e^{t\varepsilon})$ under $\varepsilon \sim N(0,\sigma^2)$

$$y_j = w^T\phi(x_j) + \epsilon$$

$$\underset{\mathcal{N}(0,\sigma^2)}{E}\left[e^{(w^T\phi(x_j)+\epsilon)t}\right] \qquad \underset{\mathcal{N}(0,\sigma^2)}{E}\left[e^{\epsilon t}\right]$$

$$= E_{\mathcal{N}(0,\sigma^2)}\left[e^{w^T\phi(x_j)t}\, e^{\epsilon t}\right] = e^{w^T\phi(x_j)t}\, \underbrace{E_{\mathcal{N}(0,\sigma^2)}\left[e^{\epsilon t}\right]}_{e^{\sigma^2 t^2/2}}$$

$$\mathcal{N}(w^T\phi(x_j), \sigma^2) \xleftarrow{\text{Given 1-1 mapping}} e^{w^T\phi(x_j)t + \sigma^2 t^2/2}$$

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

Sanity check!

$$P(y_j \mid x_j, w) = \mathcal{N}(w^T\phi(x_j), \sigma^2)$$

Need to estimate "$\underbrace{\text{the most representative}}_{\text{most likely}}$ $\underline{w}$" for given $D = \{(x_1,y_1),(x_2,y_2) \ldots (x_m,y_m)\}$

$$L(w \mid D) = \Pr((x_1,y_1),(x_2,y_2) \ldots (x_m,y_m) \,;\, w, \sigma^2, \phi(\cdot))$$

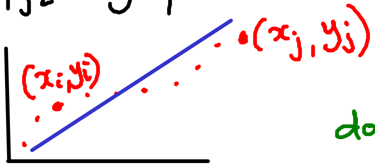
Estimating **w**: Maximum Likelihood

- 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*:

$L(w|D)$

$\Pr(\mathcal{D}|\mathbf{w}) = \Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) =$

$\Pr(y_1, y_2 \ldots y_m | x_1, x_2 \ldots x_m, \phi(\cdot), w, \sigma^2) = \prod_i \Pr(y_i|x_i$



The $(x_1,y_1) \ldots (x_m,y_m)$ collectively influence the fit (w) BUT

Given a w, prediction y_j for x_j does NOT influence y_i for x_i

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)

Tut 1 problem:

w: $x^* = \arg\max_x \Omega(x)$ $L(w|D)$

log

Let r be a monontonically increasing fn

Then: $\arg\max_x r(\Omega(x)) = x^*$ (Claim)

Optimization Trick

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

  For a fixed σ^2
  $\hat{\mathbf{w}_{ML}} =$

$$\log\left(\prod_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w^T\phi(x_j)-y_j)^2}{2\sigma^2}\right)\right) = \log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \exp\left(-\sum_j \frac{(w^T\phi(x_j)-y_j)^2}{2\sigma^2}\right)\right)$$

$$\Sigma$$

$$= -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}(w^T\phi(x_j)-y_j)^2$$

Optimization Trick

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

- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) = -\cancel{\frac{m}{2} ln(2\pi\sigma^2)} - \underbrace{\frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w^T}\phi(\mathbf{x_j}) - \mathbf{y_j})^2}$

  independent of w

  minimize over negated component

  For a fixed σ^2

  $$\hat{\mathbf{w}_{ML}} = \text{argmax } LL(y_1...y_m|\mathbf{x}_1 \ldots \mathbf{x}_m, \mathbf{w}, \sigma^2)$$

  $$= \underset{w}{\text{argmin}} \underbrace{\sum_{j=1}^{m}(\mathbf{w}^T\phi(\mathbf{x}_j) - y_j)^2}$$

- Note that this is same as the Least square solution!!

  ↳ with additional power to predict $Pr(y_j \mid x_j, \hat{w}_{ML}, \sigma^2)$

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

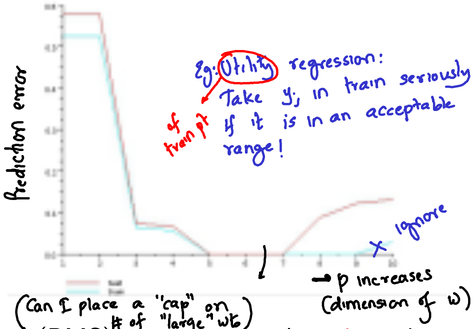


Figure 2: Root Mean Squared (RMS) errors on sample train and test datasets as a function of the degree t of the polynomial being fit

- Too many bends (t=9 onwards) in curve $\equiv$ high values of some $w_i's$. 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

Combining $\epsilon \sim N(0, \sigma^2)$ with $w_i \sim N(0, \frac{1}{\lambda})$

- 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 **w** using a prior distribution and use the posterior over **w** as the result
- Intuitive Prior: Components of **w** should not become too large!
- Next: Illustration of Bayesian Estimation on a simple Coin-tossing example

Hint from ϵ: $w_i \sim N(0, \frac{1}{\lambda})$

Implicitly putting a cap on $[\underbrace{\quad}_{w}]$ by restricting 99.7% $w_i \in [\pm 3/\sqrt{\lambda}]$

Ideally: $\underbrace{\|w\|_0}_{\text{\# of non-zero components}} \leq \theta$

Limitation: No probabilistic interpretation

Good news: An efficient algo published in 2015 on minimizing error s.t $\|w\|_0 \leq \theta$

[On Iterative Methods for Hard Thresholding, Prateek Jain, MSR I]