

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

Recap: Linear Regression is not Naively Linear

For ML, it is not just line fitting

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

$$\Phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \dots & \phi_p(\mathbf{x}_1) \\ \vdots & \vdots & & \vdots \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & \dots & \phi_n(\mathbf{x}_m) \end{bmatrix} \quad (1)$$

eg: $\phi_1(x) = \alpha_0$ $\phi_2(x) = x$... $\Rightarrow f(x) = \text{poly of degree } d$
 $\phi_1(x_1, x_2) = \alpha_0$ $\phi_{10}(x_1, x_2) = x_1$ $\phi_{01}(x_1, x_2) = x_2$...
 $\phi_{11}(x_1, x_2) = x_1 x_2$... $\phi_{pq}(x_1, x_2) = x_1^p x_2^q$

o/v examples: Radial basis fns, fourier, wavelets, formulae

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

eqs = m
features = p
(sometimes I use
features = n)

$$\Phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \dots & \phi_p(\mathbf{x}_1) \\ \vdots & \vdots & & \vdots \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & \dots & \phi_n(\mathbf{x}_m) \end{bmatrix} \quad (1)$$

- Least Squares error and corresponding estimates:

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

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

opened up square
dot product

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 $\Phi \Rightarrow$

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

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

More in tutorial

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 A posteriori 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

→ Alternative error fns ↓

Probabilistic Modeling of Linear Regression

Q: Why is Y not a deterministic fn of $w^T \phi(x) + b$?

- 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 :

reasonably behaved $Y = w^T \phi(x) + \varepsilon$
 ε noise

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

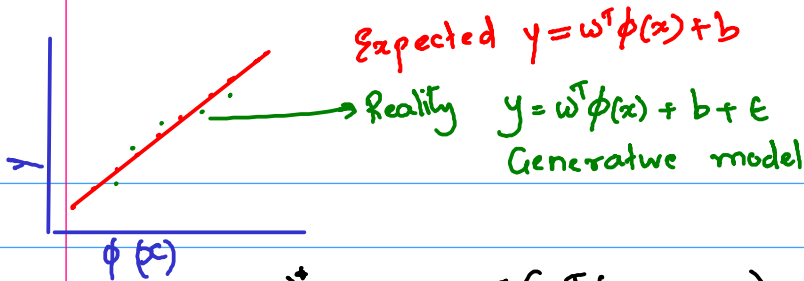
$H(p)$

- Motivation: $\mathcal{N}(\mu, \sigma^2)$, has maximum entropy among all real-valued distributions with a specified variance σ^2
- 3 - σ 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 .

$$H(p) = \int_{\mathcal{V}} -p(v) \log_2(p(v)) dv = E_p(-\log_2 p(v))$$

Noise has probabilistic semantics

Entropy is more from optimization persp.



$$\omega^* = \underset{\omega}{\operatorname{argmin}} E(\omega^T \phi(x) + b, y)$$

is more of a process of explaining the Data!

Q: Is $f(x) = \omega^T \phi(x) + b \equiv \hat{f}(x) = \hat{\omega}^T \hat{\phi}(x)$

Ans: Yes \longrightarrow

(Push b into $\hat{\omega}$ &
into $\hat{\phi}$)

with $\hat{\omega} = [\omega \ b]$

$$\hat{\phi} = [\phi \ 1]$$

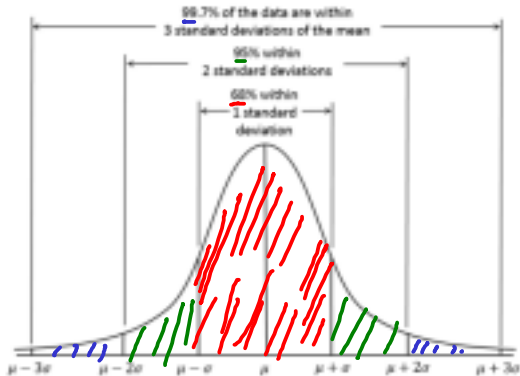
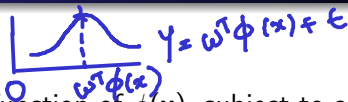


Figure 1: 3 – σ rule: About 68% of values drawn from $\mathcal{N}(\mu, \sigma^2)$ are within one standard deviation σ away from the mean μ ; about 95% of the values lie within 2σ ; and about 99.7% are within 3σ . Source: https://en.wikipedia.org/wiki/Normal_distribution

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: $E[Y(\mathbf{w}, \mathbf{x}_j)] = \mathbf{w}^T \phi(\mathbf{x}_j)$, $V(Y) = \sigma^2$

mean in Y 's shifted by $\mathbf{w}^T \phi(\mathbf{x}_j)$ with respect to ε

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

Estimating \mathbf{w} : Maximum Likelihood

- If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$ where $\mathbf{w}, \phi(\mathbf{x}) \in \mathbb{R}^m$ then, given dataset \mathcal{D} , find the most likely $\mathbf{w}_{ML}^{\hat{}}$

- Recall: $\Pr(y_j | \mathbf{x}_j, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - \mathbf{w}^T \phi(\mathbf{x}_j))^2}{2\sigma^2}\right) = \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}_j), \sigma^2)$

- From *Probability of data to likelihood of parameters*:

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

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

How likely
is the data, given \mathbf{w}

$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \end{bmatrix}$

Observations $(x_1, y_1), (x_2, y_2)$
 $\dots (x_m, y_m)$

are all independent &
have identical distribution
(i.i.d distributed $\langle x_i, y_i \rangle$)

Estimating \mathbf{w} : Maximum Likelihood

- If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$ where \mathbf{w} , $\phi(\mathbf{x}) \in \mathbf{R}^m$ then, given dataset \mathcal{D} , find the most likely \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

$$\hat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}} \Pr(\mathcal{D} | \mathbf{w}) = \Pr(\mathbf{y} | \mathbf{x}, \mathbf{w}) = L(\mathbf{w} | \mathcal{D})$$

to be found ←
fixed or known ↓

Optimization Trick

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

$$\begin{aligned} & \log \left(\prod_{j=1}^m \left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{(\omega^T \phi(x_j) - y_j)^2}{2\sigma^2}} \right) \\ &= \log \left[\left[\frac{1}{\sigma\sqrt{2\pi}} \right]^m e^{-\sum_{j=1}^m \frac{(\omega^T \phi(x_j) - y_j)^2}{2\sigma^2}} \right] \\ &= -m \log(\sigma\sqrt{2\pi}) - \sum_{j=1}^m \frac{(\omega^T \phi(x_j) - y_j)^2}{2\sigma^2} \end{aligned}$$

Optimization Trick

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

- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$ [LL is called log-likelihood]

$$-\frac{m}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^m (\mathbf{w}^T \phi(\mathbf{x}_j) - y_j)^2$$

For a fixed σ^2

$$\mathbf{w}_{ML}^{\hat{}} = \underset{\mathbf{w}}{\operatorname{argmax}} LL(\mathbf{w}|\mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmax}} -\frac{1}{2\sigma^2} \sum_{j=1}^m (\mathbf{w}^T \phi(\mathbf{x}_j) - y_j)^2$$

Optimization Trick

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

- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$

$$-\frac{m}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^m (\mathbf{w}^T \phi(\mathbf{x}_j) - y_j)^2$$

For a fixed σ^2

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

$$= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{j=1}^m (\mathbf{w}^T \phi(\mathbf{x}_j) - y_j)^2 = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^m (\mathbf{w}^T \phi(\mathbf{x}_i) - y_i)^2$$

- 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 A posteriori 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

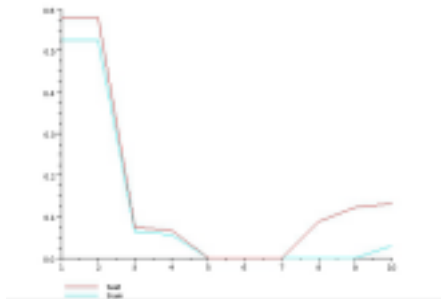


Figure 2: Root Mean Squared (RMS) errors on sample **train** and **test** datasets as a function of the degree t of the polynomial being fit

- 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

$X^0 *$	0.13252679175596802
$X^1 *$	6.836159339696569
$X^2 *$	-10.198794083500966
$X^3 *$	8.298738913209064
$X^4 *$	-3.766949862252123
$X^5 *$	1.0274981119277349
$X^6 *$	-0.17218031550131038
$X^7 *$	0.017340835860554016
$X^8 *$	-9.623065771393043E-4
$X^9 *$	2.2595409656184083E-5

w vector with fewer pts

$$\|w\| < \|\tilde{w}\|$$

$X^0 *$	-1.4218758581602278
$X^1 *$	14.756472312089675
$X^2 *$	-24.299789484296475
$X^3 *$	20.63606795357865
$X^4 *$	-9.934453145766518
$X^5 *$	2.8975181063446613

\tilde{w} vector with more "distracting" points

Note: ① Test data cannot be used to decide that we are overfitting



② Use validation/hold-out data as proxy to test

③ Use $\|w\|$ or some measure of deviation of \hat{w}_{ML} from an "expected prior" behaviour as a proxy to detect overfitting



Test

Motivation for ③ is training data is

EXPENSIVE

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