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)!

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

Eg: $\phi_1(x) = x_0$ $\phi_2(x) = x$ $\ldots$ $\Rightarrow f(x) =$ poly of degree d

$\phi_1(x_1, x_2) = x_0$ $\phi_{10}(x_1, x_2) = x_1$ $\phi_{01}(x_1, x_2) = x_2$ $\ldots$

$\phi_{11}(x_1, x_2) = x_1 x_2$ $\ldots$ $\phi_{pq}(x_1, x_2) = x_1^p x_2^q$

o/w examples: Radial basis fns, fourier, wavelets, formulae in logic

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)!

#egs = m
#features = p
(sometimes I use
features = n)

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

- Least Squares error and corresponding estimates:

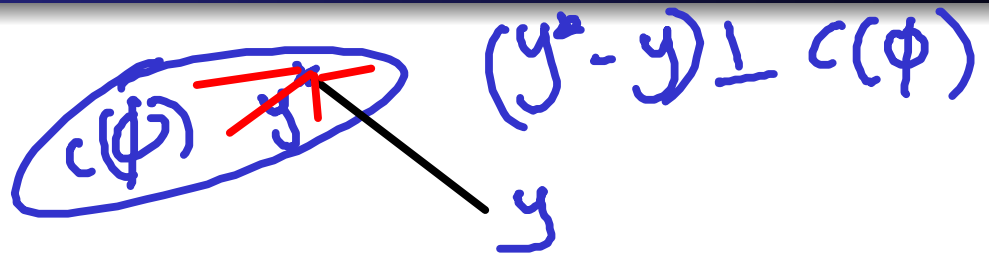
$$E^* = \min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \left(\mathbf{w^T}\phi^{\mathbf{T}}\phi\mathbf{w} - \mathbf{2y^T}\phi\mathbf{w} + \mathbf{y^T y} \right) \tag{2}$$

opened up square

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

→ opened up 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 $\phi \Rightarrow$

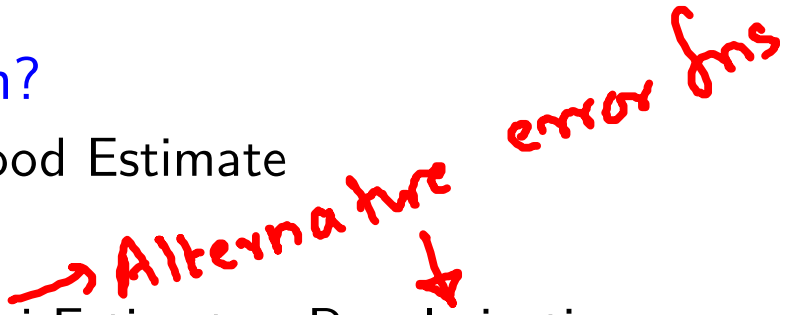
$$\mathbf{w} = (\phi^{\mathbf{T}}\phi)^{-\mathbf{1}}\phi\mathbf{y} \tag{4}$$

- 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 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

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

Noise has probabilistic semantics

Error is more flexible optimization perspective

- 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 ε noise

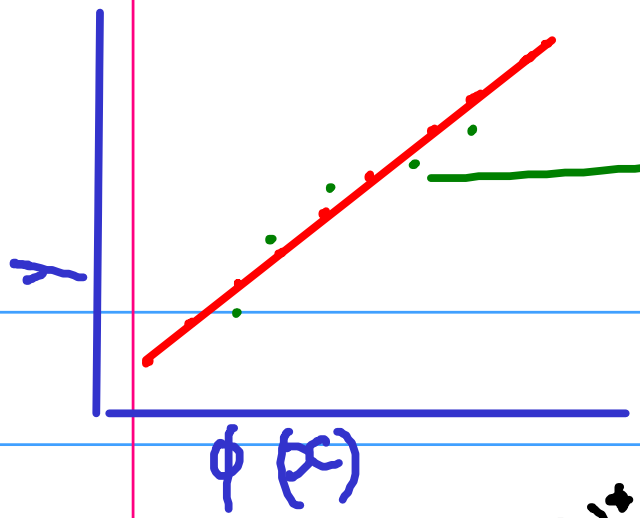
$$Y = w^T\phi(x) + \varepsilon$$

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

H(p)

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

$$H(p) = \int_v -p(v)\log_2(p(v))\,dv = E_p(-\log_2 p(v))$$



Expected $y = w^T \phi(x) + b$

Reality $y = w^T \phi(x) + b + \epsilon$

Generative model

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

is more of a process of explaining the Data!

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

Ans: Yes ⟶ with $\hat{w} = [w\ \ b]$

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

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

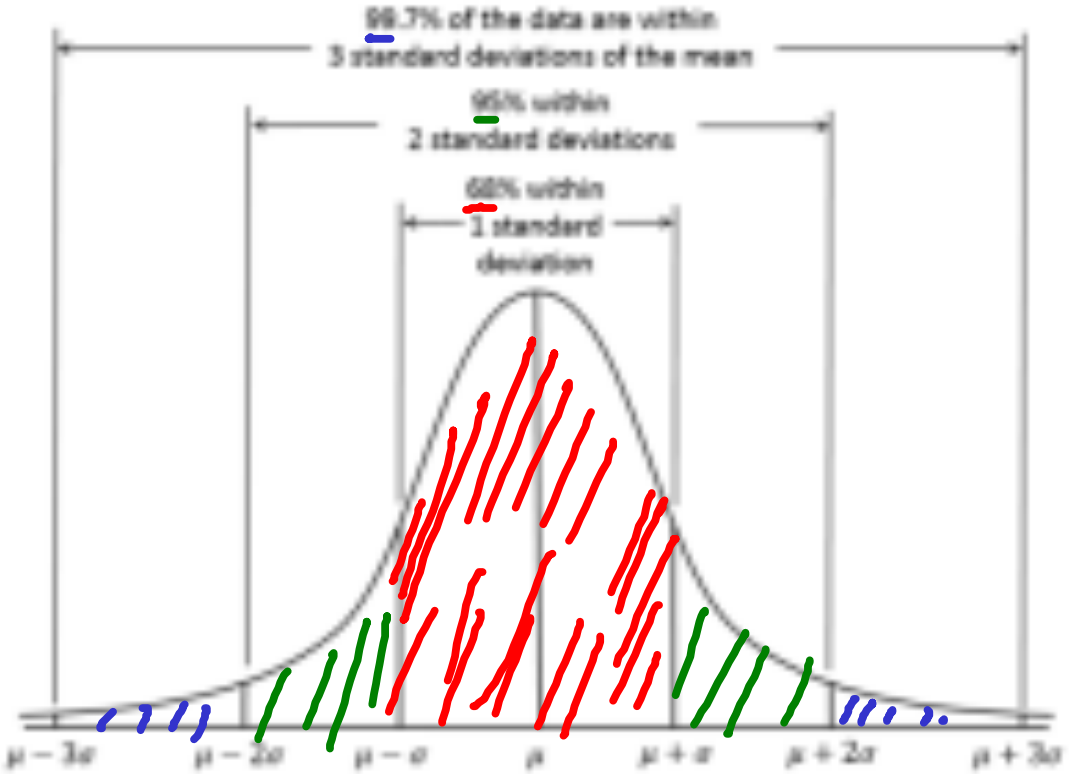
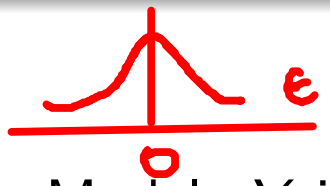
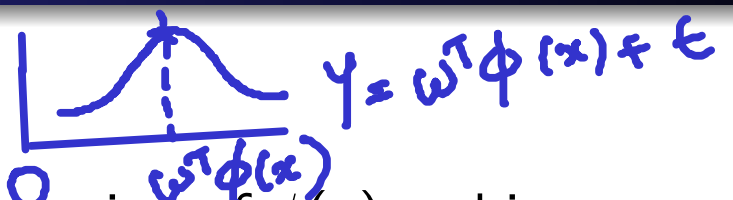


Figure 1: $3-\sigma$ rule: About 68% of values drawn from $\mathcal{N}(\mu, \sigma^2)$ are within one standard deviation σ away from the mean μ; about 95% of the values lie within 2σ; and about 99.7% are within 3σ. Source: `https://en.wikipedia.org/wiki/Normal_distribution`

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:

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

mean shifted in y_i is with respect to ε by $w^T\phi(x_i)$

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

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

Estimating **w**: Maximum Likelihood

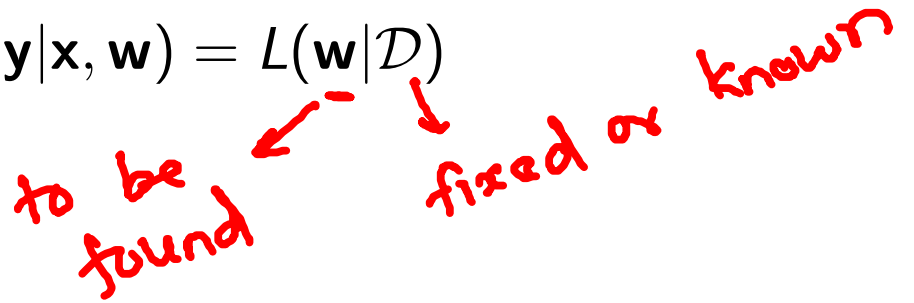
- If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$ where $\mathbf{w}$, $\phi(\mathbf{x}) \in \mathbf{R}^m$ then, given dataset $\mathcal{D}$, find the most likely $\hat{\mathbf{w}}_{ML}$
- Recall: $\Pr(y_j|\mathbf{x}_j, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}}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

How likely is the data, given w $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} \to \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}$

Observations $(x_1, y_1), (x_2, y_2) \ldots (x_m, y_m)$ 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 $y = \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x}) + \epsilon$ where $\mathbf{w},\ \phi(\mathbf{x}) \in \mathbf{R^m}$ 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 $\hat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\text{argmax}} \Pr(\mathcal{D}|\mathbf{w}) = \Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) = L(\mathbf{w}|\mathcal{D})$



Optimization Trick

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

$$\log_e\left(\prod_{j=1}^{m}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)e^{-\left(\frac{(w^T\phi(x_j)-y_j)^2}{2\sigma^2}\right)}\right)$$

$$= \log_e\left[\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^m e^{-\left(\sum_{j=1}^{m}\frac{(w^T\phi(x_j)-y_j)^2}{2\sigma^2}\right)}\right]$$

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

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as log)
- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$ [LL is called log-likelihood]

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

For a fixed σ^2

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

Optimization Trick

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

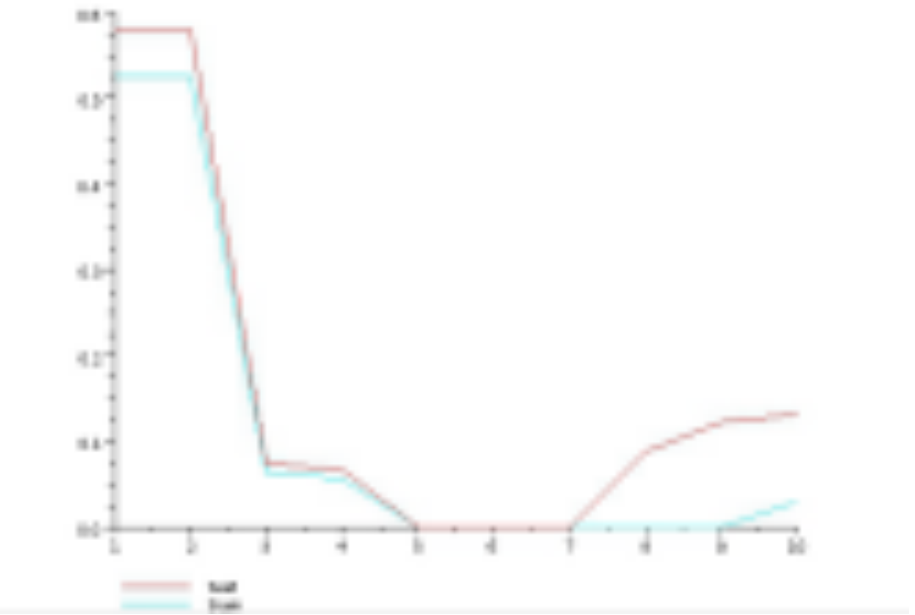


Figure 2: Root Mean Squared (RMS) errors on sample train and test datasets as a function of the degree t of the polynomial being fit

- 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

X^0 * 0.1325267917559680 2
X^1 * 6.836159339696569
X^2 * -10.19879408350096 6
X^3 * 8.298738913209064
X^4 * -3.766949862252123
X^5 * 1.027498111927734 9
X^6 * -0.172180315501310 38
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

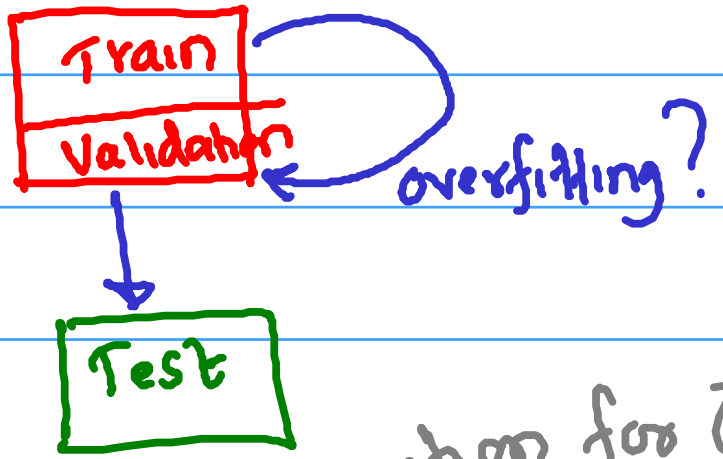
$\tilde{w}$ vector with more "distracting" points

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

options

② Use data

validation / hold-out as proxy to test data



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

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 **w** using a prior distribution and use the posterior over **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