CS545: Linear Models (Nonlinear Inputs; Probabilistic)

Chuck Anderson

Department of Computer Science
Colorado State University

Fall, 2009

Outline

Linear in Parameters, Nonlinear in Inputs

- Linear dependence on parameters can be maintained if we substitute for the x_i some nonlinear functions of $\mathbf{x}$.

$$
\begin{aligned}
y(\mathbf{x}, \mathbf{w}) &= w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \cdots + w_{M-1}\phi_{M-1}(\mathbf{x}) \\
&= \mathbf{w}^T\phi(\mathbf{x}), \text{ where } \phi_0(\mathbf{x}) = 1
\end{aligned}
$$

Linear in Parameters, Nonlinear in Inputs

- Linear dependence on parameters can be maintained if we substitute for the x_i some nonlinear functions of $\mathbf{x}$.

$$
\begin{aligned}
y(\mathbf{x},\mathbf{w}) &= w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \cdots + w_{M-1}\phi_{M-1}(\mathbf{x}) \\
&= \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), \text{ where } \phi_0(\mathbf{x}) = 1
\end{aligned}
$$

- The functions ϕ_i could be any functions, but typical examples are

$$
\begin{aligned}
\phi_i(\mathbf{x}) &= x_i \text{ makes } y(\mathbf{x},\mathbf{w}) \text{ linear in } \mathbf{x} \\
&= x_j^k \text{ or } x_j x_k \\
&= e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\
&= \frac{1}{1+e^{-\mathbf{a}^T\mathbf{x}}} \\
&= \tanh(\mathbf{a}^T\mathbf{x})
\end{aligned}
$$

Example: Noisy Sine

- $f(x) = \sin(3x) + \mathcal{N}(0, 0.2)$ where $x = [0, 1]$

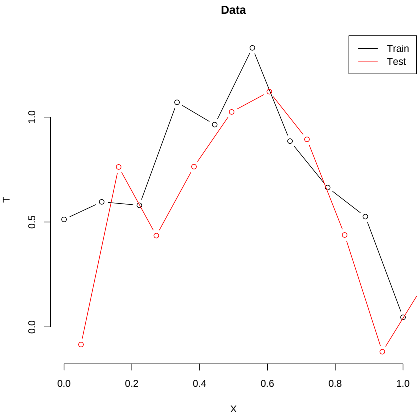
Example: Noisy Sine

- $f(x) = \sin(3x) + \mathcal{N}(0, 0.2)$ where $x = [0, 1]$

Example: Noisy Sine

- $f(x) = \sin(3x) + \mathcal{N}(0, 0.2)$ where $x = [0, 1]$

```
X <- matrix(seq(0,1,len=10))
f <- function(X) matrix(sin(3*X) + rnorm(length(X),0,0.2))
T <- f(X)
Xtest <- X +0.05
Ttest <- f(Xtest)
```



Polynomial Functions

- Linear model won't work well. Need to use some nonlinear functions, $\phi(x)$ of the input. Which ones?

Polynomial Functions

- Linear model won't work well. Need to use some nonlinear functions, $\phi(x)$ of the input. Which ones?
- Let's try polynomials.

Polynomial Functions

- Linear model won't work well. Need to use some nonlinear functions, $\phi(x)$ of the input. Which ones?
- Let's try polynomials.

Polynomial Functions

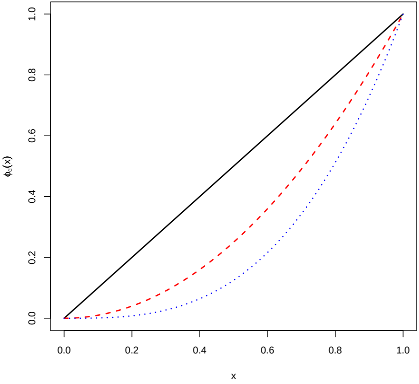
- Linear model won't work well. Need to use some nonlinear functions, $\phi(x)$ of the input. Which ones?
- Let's try polynomials.

$$
\begin{aligned}
\phi_1(x) &= x \\
\phi_2(x) &= x^2 \\
\phi_3(x) &= x^3 \\
&\vdots \\
\phi_d(x) &= x^d
\end{aligned}
$$

Polynomial Functions

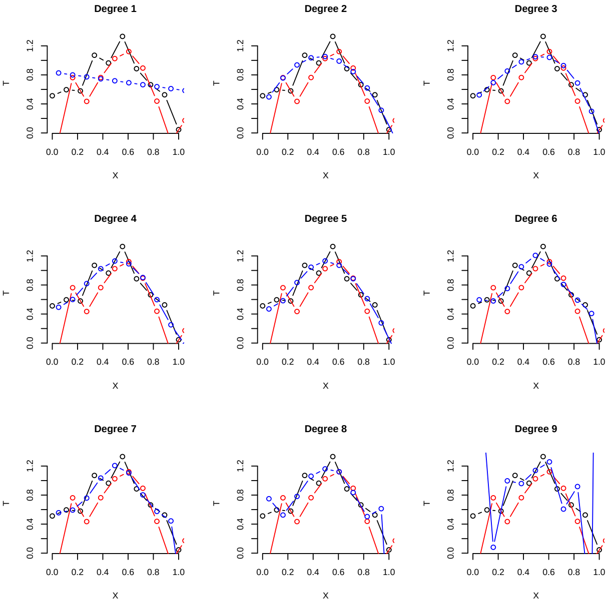
- Linear model won't work well. Need to use some nonlinear functions, $\phi(x)$ of the input. Which ones?
- Let's try polynomials.

$$\phi_1(x) = x$$
$$\phi_2(x) = x^2$$
$$\phi_3(x) = x^3$$
$$\vdots$$
$$\phi_d(x) = x^d$$

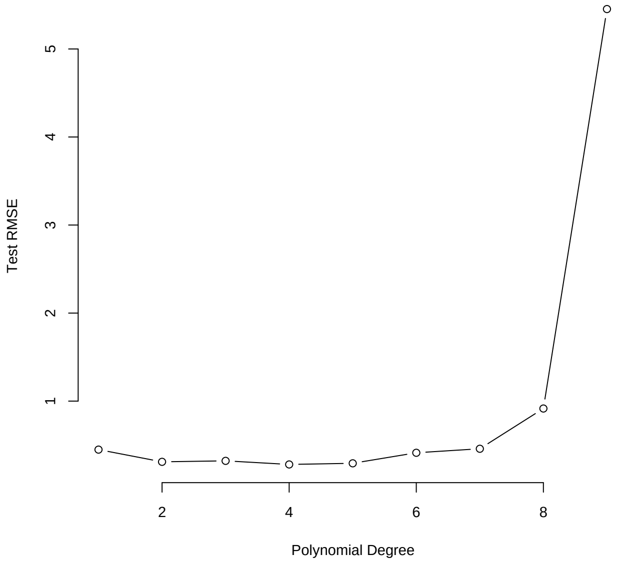


Results for Different Maximum Degrees *d*

black - training data red - testing data blue - model on testing data



Results for Different Maximum Degrees d



Radial Basis Functions

- Another popular choice is radial basis functions, or RBFs.

Radial Basis Functions

- Another popular choice is radial basis functions, or RBFs.

Radial Basis Functions

- Another popular choice is radial basis functions, or RBFs.

$$
\begin{aligned}
\phi_1(x) &= e^{-\frac{(x-0)^2}{0.1}} \\
\phi_2(x) &= e^{-\frac{(x-0.2)^2}{0.1}} \\
\phi_3(x) &= e^{-\frac{(x-0.4)^2}{0.1}} \\
&\vdots
\end{aligned}
$$

Radial Basis Functions

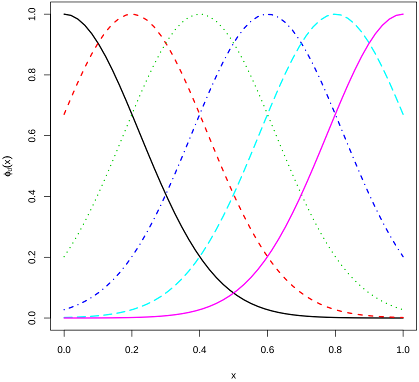
- Another popular choice is radial basis functions, or RBFs.

$$\phi_1(x) = e^{-\frac{(x-0)^2}{0.1}}$$

$$\phi_2(x) = e^{-\frac{(x-0.2)^2}{0.1}}$$

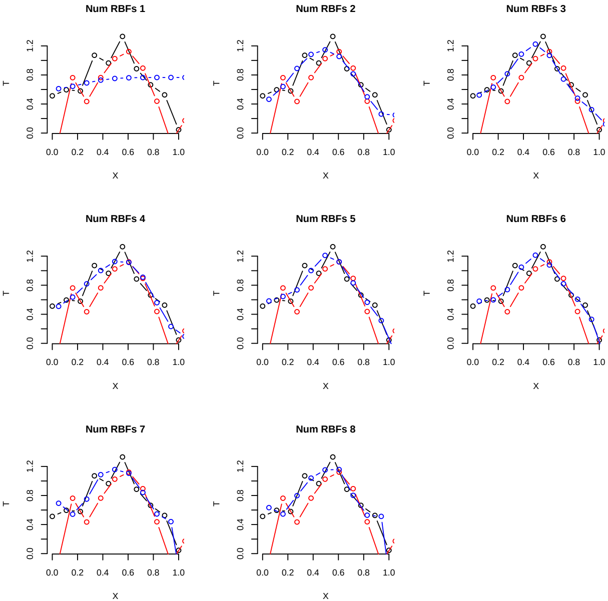
$$\phi_3(x) = e^{-\frac{(x-0.4)^2}{0.1}}$$

$$\vdots$$

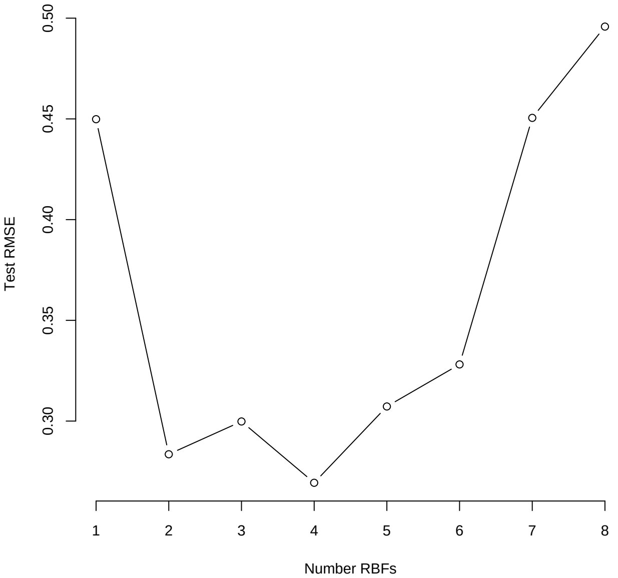


Results for Different Numbers of RBFs

black - training data red - testing data blue - model on testing data

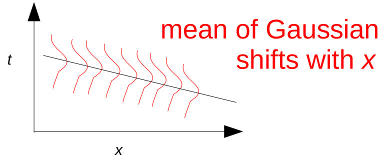


Results for Different Numbers of RBFs



Outline

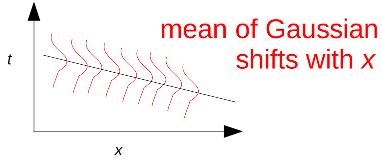
Linear Model as Probabilistic Model



- To enter the probabilistic world, let's say our model $y(\mathbf{x}_n, \mathbf{w})$ predicts t_n with an error that is modeled as a Gaussian random variable with precision β.

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon$$
$$p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) = \mathcal{N}(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1})$$

Linear Model as Probabilistic Model



- To enter the probabilistic world, let's say our model $y(\mathbf{x}_n, \mathbf{w})$ predicts t_n with an error that is modeled as a Gaussian random variable with precision β.

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon$$

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) = \mathcal{N}(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1})$$

- Our prediction of t_n for a given sample $\mathbf{x}_n$ is now not a single value, but a distribution over possible values—its expected value conditioned on $\mathbf{x}_n$.

$$E[t_n|\mathbf{x}_n] = \int t \; p(t_n|\mathbf{x}_n) dt = y(\mathbf{x}_n, \mathbf{w})$$

- The likelihood for all data samples is

$$
\begin{aligned}
p(T|\mathbf{X}, \mathbf{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}(t_n | y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}) \\
&= \prod_{n=1}^{N} \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})
\end{aligned}
$$

where $\phi(\mathbf{x}_n)$ is a vector of the values of all of the nonlinear functions (sometimes called basis functions) applied to the sample $\mathbf{x}_n$.

Fitting Model by Maximum Likelihood

- Taking the logarithm of the likelihood we get

$$\ln p(T|\mathbf{X}, \mathbf{w}, \beta) = -\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 - \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

- Then take derivative, actually gradient, with respect to $\mathbf{w}$. (Just like minimizing squared error before we entered the probabilistic realm!)

Fitting Model by Maximum Likelihood

- Taking the logarithm of the likelihood we get

$$\ln p(T|\mathbf{X},\mathbf{w},\beta) = -\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 - \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

- Then take derivative, actually gradient, with respect to **w**. (Just like minimizing squared error before we entered the probabilistic realm!)

$$\nabla_{\mathbf{w}} \ln p(T|\mathbf{X},\mathbf{w},\beta) = \sum_{n=1}^{N}(t_n - \mathbf{w}^T\phi(\mathbf{x}_n))\phi(\mathbf{x}_n)^T$$

- Setting this equal to zero we can solve for **w**.

$$0 = \sum_{n=1}^{N} t_n\phi(\mathbf{x}_n)^T - \mathbf{w}^T\left(\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^T\right)$$

Fitting Model by Maximum Likelihood

- These sums can be expressed as matrix operations if we define

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{D-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{D-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{D-1}(\mathbf{x}_N) \end{pmatrix}$$

  Now the above equation becomes the following one and the solution for **w** continues.

$$\begin{aligned} 0 &= \mathbf{\Phi}^T T - \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} \\ \mathbf{\Phi}^T T &= \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} \\ \mathbf{w} &= (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T T \end{aligned}$$

- So, fitting a probabilistic model defined for Gaussian distribution of fixed precision and mean as a linear function of inputs, gives same solution as the non-probabilistic least squares solution.

Maximum a Posterior

- Alternative to maximizing the likelihood of the data is to maximize the posterior distribution of $\mathbf{w} = p(\mathbf{w}|\mathbf{X}, \mathbf{T}, \text{parameters})$, where parameters represents basis functions and distribution parameters our model will be based on.

Maximum a Posterior

- Alternative to maximizing the likelihood of the data is to maximize the posterior distribution of $\mathbf{w} = p(\mathbf{w}|\mathbf{X}, \mathbf{T}, \text{parameters})$, where parameters represents basis functions and distribution parameters our model will be based on.
- Called the MAP solution (versus the ML, or maximum likelihood solution).

Maximum a Posterior

- Alternative to maximizing the likelihood of the data is to maximize the posterior distribution of $\mathbf{w} = p(\mathbf{w}|\mathbf{X}, \mathbf{T}, \text{parameters})$, where parameters represents basis functions and distribution parameters our model will be based on.
- Called the MAP solution (versus the ML, or maximum likelihood solution).
- Using Bayes Theorem, we know (leaving out the parameters)

$$p(\mathbf{w}|\mathbf{X}, \mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{T})}$$

Maximum a Posterior

- Alternative to maximizing the likelihood of the data is to maximize the posterior distribution of $\mathbf{w} = p(\mathbf{w}|\mathbf{X}, \mathbf{T}, \text{parameters})$, where parameters represents basis functions and distribution parameters our model will be based on.
- Called the MAP solution (versus the ML, or maximum likelihood solution).
- Using Bayes Theorem, we know (leaving out the parameters)

$$p(\mathbf{w}|\mathbf{X}, \mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{T})}$$

- We will be maximizing this by finding best $\mathbf{w}$ which does not affect the denominator, so can work just with the numerator.

Choosing $p(\mathbf{w})$

- Must choose a prior distribution for $\mathbf{w}$.

Choosing $p(\mathbf{w})$

- Must choose a prior distribution for $\mathbf{w}$.
- What value of $\mathbf{w}$ would we prefer if no training data exists?

Choosing $p(\mathbf{w})$

- Must choose a prior distribution for $\mathbf{w}$.
- What value of $\mathbf{w}$ would we prefer if no training data exists?
- $\mathbf{w} = 0$

Choosing $p(\mathbf{w})$

- Must choose a prior distribution for $\mathbf{w}$.
- What value of $\mathbf{w}$ would we prefer if no training data exists?
- $\mathbf{w} = 0$
- With small amount of data, let's still strongly bias it towards something close to zero.

Choosing $p(\mathbf{w})$

- Must choose a prior distribution for $\mathbf{w}$.
- What value of $\mathbf{w}$ would we prefer if no training data exists?
- $\mathbf{w} = 0$
- With small amount of data, let's still strongly bias it towards something close to zero.
- Gaussian distribution will do this, and it is mathematically convenient. Use mean of zero and precision that we choose empirically to control how strongly we want to force $\mathbf{w}$ to stay close to zero.

$$p(\mathbf{w}|\alpha) = \mathcal{N}(w|\alpha^{-1}I)$$

Choosing $p(\mathbf{w})$

- Now, using our previous probabilistic model for $p(\mathbf{T}|\mathbf{X},\mathbf{w},\beta)$, we have

$$p(\mathbf{w}|\mathbf{X},\mathbf{T},\alpha,\beta) \propto p(\mathbf{T}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$$

$$p(\mathbf{T}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\alpha) = \prod_{n=1}^{N} \mathcal{N}(t_n|\phi(\mathbf{x}_n)\mathbf{w},\beta^{-1})\mathcal{N}(\mathbf{w}|\alpha^{-1}I)$$

Choosing $p(\mathbf{w})$

- Now, using our previous probabilistic model for $p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \beta)$, we have

$$p(\mathbf{w}|\mathbf{X}, \mathbf{T}, \alpha, \beta) \propto p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

$$p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) = \prod_{n=1}^{N} \mathcal{N}(t_n|\phi(\mathbf{x}_n)\mathbf{w}, \beta^{-1})\mathcal{N}(\mathbf{w}|\alpha^{-1}I)$$

- To maximize this, take the logarithm.

$$\sum_{n=1}^{N} \ln \mathcal{N}(t_n|\phi(\mathbf{x}_n)\mathbf{w}, \beta^{-1}) + \ln \mathcal{N}(\mathbf{w}|\alpha^{-1}I)$$

$$= -\frac{1}{2}\beta \sum_{n=1}^{N}(t_n - \phi(\mathbf{x}_n)\mathbf{w})^2 + \frac{1}{2}N \ln \beta - \frac{1}{2}N \ln(2\pi)$$

$$- \frac{1}{2}\mathbf{w}^T \alpha I \mathbf{w} + \frac{1}{2}\alpha I - \frac{1}{2}\ln(2\pi)$$

Choosing $p(\mathbf{w})$

- Now, using our previous probabilistic model for $p(\mathbf{T}|\mathbf{X},\mathbf{w},\beta)$, we have

$$p(\mathbf{w}|\mathbf{X},\mathbf{T},\alpha,\beta) \propto p(\mathbf{T}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$$

$$p(\mathbf{T}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\alpha) = \prod_{n=1}^{N} \mathcal{N}(t_n|\phi(\mathbf{x}_n)\mathbf{w},\beta^{-1})\mathcal{N}(\mathbf{w}|\alpha^{-1}I)$$

- To maximize this, take the logarithm.

$$\begin{aligned}
&\sum_{n=1}^{N} \ln \mathcal{N}(t_n|\phi(\mathbf{x}_n)\mathbf{w},\beta^{-1}) + \ln \mathcal{N}(\mathbf{w}|\alpha^{-1}I) \\
&\quad = -\frac{1}{2}\beta \sum_{n=1}^{N} (t_n - \phi(\mathbf{x}_n)\mathbf{w})^2 + \frac{1}{2}N\ln\beta - \frac{1}{2}N\ln(2\pi) \\
&\qquad - \frac{1}{2}\mathbf{w}^T\alpha I\mathbf{w} + \frac{1}{2}\alpha I - \frac{1}{2}\ln(2\pi)
\end{aligned}$$

- Now take the gradient of this with respect to $\mathbf{w}$.

- The gradient of this with respect to $\mathbf{w}$ is

$$\beta \sum_{n=1}^{N} (t_n - \phi(\mathbf{x}_n)\mathbf{w})\phi(\mathbf{x}_n)^T - \alpha I \mathbf{w}$$

- The gradient of this with respect to **w** is

$$\beta \sum_{n=1}^{N} (t_n - \phi(\mathbf{x}_n)\mathbf{w})\phi(\mathbf{x}_n)^T - \alpha I \mathbf{w}$$

- Setting this equal to zero and rearranging a bit we get

$$\begin{aligned} 0 &= \beta \left(\sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^T - \mathbf{w}^T \sum_{n=1}^{N} \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^T \right) - \alpha I \mathbf{w} \\ &= \beta(\mathbf{\Phi}^T \mathbf{T} - \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w}) - \alpha I \mathbf{w} \end{aligned}$$

- The gradient of this with respect to $\mathbf{w}$ is

$$\beta \sum_{n=1}^{N} (t_n - \phi(\mathbf{x}_n)\mathbf{w})\phi(\mathbf{x}_n)^T - \alpha I \mathbf{w}$$

- Setting this equal to zero and rearranging a bit we get

$$\begin{aligned} 0 &= \beta \left(\sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^T - \mathbf{w}^T \sum_{n=1}^{N} \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^T \right) - \alpha I \mathbf{w} \\ &= \beta(\mathbf{\Phi}^T \mathbf{T} - \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w}) - \alpha I \mathbf{w} \end{aligned}$$

- Solving for $\mathbf{w}$ results in

$$\begin{aligned} 0 &= \beta(\mathbf{\Phi}^T \mathbf{T} - \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w}) - \alpha I \mathbf{w} \\ 0 &= \mathbf{\Phi}^T \mathbf{T} - \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} - \frac{\alpha}{\beta} I \mathbf{w} \\ \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} + \frac{\alpha}{\beta} I \mathbf{w} &= \mathbf{\Phi}^T \mathbf{T} \\ (\mathbf{\Phi}^T \mathbf{\Phi} + \frac{\alpha}{\beta} I)\mathbf{w} &= \mathbf{\Phi}^T \mathbf{T} \\ \mathbf{w} &= (\mathbf{\Phi}^T \mathbf{\Phi} + \frac{\alpha}{\beta} I)^{-1} \mathbf{\Phi}^T \mathbf{T} \end{aligned}$$

- Does this look familiar? Let's call the solution $\mathbf{w}_{\text{MAP}}$.

$$\mathbf{w}_{\text{MAP}} = (\mathbf{\Phi}^T\mathbf{\Phi} + \frac{\alpha}{\beta} I)^{-1}\mathbf{\Phi}^T\mathbf{T}$$

- Does this look familiar? Let's call the solution $\mathbf{w}_{\text{MAP}}$.

$$\mathbf{w}_{\text{MAP}} = (\mathbf{\Phi}^T \mathbf{\Phi} + \frac{\alpha}{\beta} I)^{-1} \mathbf{\Phi}^T \mathbf{T}$$

- Should remind you of our least squares solution with a weight penalty.

$$\mathbf{w}_{\text{LS}} = (\mathbf{\Phi}^T \mathbf{\Phi} + \lambda I)^{-1} \mathbf{\Phi}^T \mathbf{T}$$