Regression II

Jiayu Zhou

[1] Department of Computer Science and Engineering
Michigan State University
East Lansing, MI USA

Table of contents

Solving the Least Squares using Normal Equation

- We call $\mathbf{\Phi}^\dagger \equiv (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ pseudo-inverse (generalization of inverse to non-square matrix). See for a square invertable matrix $\mathbf{\Phi}$, we have $\mathbf{\Phi}^\dagger = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T = \mathbf{\Phi}^{-1}(\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}^T = \mathbf{\Phi}^{-1}$

Solving the Least Squares using Normal Equation

- We call $\mathbf{\Phi}^{\dagger} \equiv (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ pseudo-inverse (generalization of inverse to non-square matrix). See for a square invertable matrix $\mathbf{\Phi}$, we have $\mathbf{\Phi}^{\dagger} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T = \mathbf{\Phi}^{-1}(\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}^T = \mathbf{\Phi}^{-1}$
- When rank$(\mathbf{\Phi}) = r \leq \min(M, N)$, let $\mathbf{\Phi} = U_r\Sigma_r V_r^T$ be the economical SVD

$$\begin{aligned}
\mathbf{w} &= (V_r\Sigma_r U_r^T U_r\Sigma_r V_r^T)^{-1} V_r S_r U_r^T \mathbf{t} \\
&= V_r\Sigma_r^{-2} V_r^T V_r\Sigma_r U_r^T \mathbf{t} \\
&= V_r\Sigma_r^{-1} U_r^T \mathbf{t} = \sum_{i=1}^{r} \frac{u_i^T\mathbf{t}}{\sigma_i} v_i
\end{aligned}$$

Solving the Least Squares using Normal Equation

- We call $\mathbf{\Phi}^\dagger \equiv (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ pseudo-inverse (generalization of inverse to non-square matrix). See for a square invertable matrix $\mathbf{\Phi}$, we have $\mathbf{\Phi}^\dagger = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T = \mathbf{\Phi}^{-1}(\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}^T = \mathbf{\Phi}^{-1}$
- When rank($\mathbf{\Phi}$) $= r \leq \min(M, N)$, let $\mathbf{\Phi} = U_r\Sigma_r V_r^T$ be the economical SVD

$$\begin{aligned}
\mathbf{w} &= (V_r\Sigma_r U_r^T U_r\Sigma_r V_r^T)^{-1} V_r S_r U_r^T \mathbf{t} \\
&= V_r\Sigma_r^{-2} V_r^T V_r\Sigma_r U_r^T \mathbf{t} \\
&= V_r\Sigma_r^{-1} U_r^T \mathbf{t} = \sum_{i=1}^{r} \frac{u_i^T \mathbf{t}}{\sigma_i} v_i
\end{aligned}$$

- How about when $\sigma_i \to 0$?

Regularization

- Adding regularization to control overfitting

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{1}$$

Regularization

- Adding regularization to control overfitting

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \quad (1)$$

- Setting gradient to zero,

Regularization

- Adding regularization to control overfitting

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{1}$$

- Setting gradient to zero,

$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^T\mathbf{t} + \lambda\mathbf{w} = 0 \Rightarrow (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)\mathbf{w} - \mathbf{\Phi}^T\mathbf{t} = 0$$

$$\Rightarrow \mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)^{-1}\mathbf{\Phi}^T\mathbf{t}$$

Regularization

- Adding regularization to control overfitting
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w}-\mathbf{t}\|_2^2+\frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{1}$$
- Setting gradient to zero,
$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}-\mathbf{\Phi}^T\mathbf{t}+\lambda\mathbf{w}=0 \Rightarrow (\mathbf{\Phi}^T\mathbf{\Phi}+\lambda I)\mathbf{w}-\mathbf{\Phi}^T\mathbf{t}=0$$
$$\Rightarrow \mathbf{w}=(\mathbf{\Phi}^T\mathbf{\Phi}+\lambda I)^{-1}\mathbf{\Phi}^T\mathbf{t}$$
- Plugin the SVD $\mathbf{\Phi}=USV^T=U\begin{bmatrix}\Sigma\\0\end{bmatrix}V$:

Regularization

- Adding regularization to control overfitting

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{1}$$

- Setting gradient to zero,

$$\mathbf{\Phi}^T\mathbf{\Phi w} - \mathbf{\Phi}^T\mathbf{t} + \lambda\mathbf{w} = 0 \Rightarrow (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)\mathbf{w} - \mathbf{\Phi}^T\mathbf{t} = 0$$

$$\Rightarrow \mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)^{-1}\mathbf{\Phi}^T\mathbf{t}$$

- Plugin the SVD $\mathbf{\Phi} = USV^T = U\begin{bmatrix}\Sigma \\ 0\end{bmatrix}V$:

$$\Rightarrow \mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I)^{-1}\mathbf{\Phi}^T\mathbf{t} = (VSU^TUSV^T + V\lambda IV^T)^{-1}VSU^T\mathbf{t}$$

$$= V(\Sigma^2 + \lambda I)^{-1}SU^T\mathbf{t} = \sum_{i=1}^{M}\frac{\sigma_i u_i^T\mathbf{t}}{\sigma_i^2 + \lambda}v_i$$

Probabilistic interpretation

- Introduce a prior probability distribution over the model parameters $\mathbf{w}$. Assume the noise precision β is known.

Probabilistic interpretation

- Introduce a prior probability distribution over the model parameters $\mathbf{w}$. Assume the noise precision β is known.
- Recall that the likelihood is given by

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

Probabilistic interpretation

- Introduce a prior probability distribution over the model parameters $\mathbf{w}$. Assume the noise precision β is known.
- Recall that the likelihood is given by

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

- Beyesian Theorem that relates prior $p(\mathbf{w}|\lambda)$ to posterior $p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda)$:

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\lambda)$$

Probabilistic interpretation

- Introduce a prior probability distribution over the model parameters $\mathbf{w}$. Assume the noise precision β is known.
- Recall that the likelihood is given by

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

- Beyesian Theorem that relates prior $p(\mathbf{w}|\lambda)$ to posterior $p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda)$:

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\lambda)$$

  - Using conjugate prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{S}_0)$, we can show that Posterior $p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$ where $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{t})$, $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$.

Probabilistic interpretation

- Introduce a prior probability distribution over the model parameters $\mathbf{w}$. Assume the noise precision β is known.
- Recall that the likelihood is given by

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

- Beyesian Theorem that relates prior $p(\mathbf{w}|\lambda)$ to posterior $p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda)$:

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\lambda)$$

  - Using conjugate prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{S}_0)$, we can show that Posterior $p(\mathbf{w}|\mathbf{t},\mathbf{X},\beta,\lambda) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$ where $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{t})$, $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$.
  - This is your home work.

- When $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$, we have $\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$, $\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$

- When $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$, we have $\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$, $\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$
- The maximum a posteriori (MAP) estimation is given by

$$\ln p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \beta, \lambda) = -\frac{\beta}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{consts.}$$

$$\Rightarrow \arg\max_{\mathbf{w}} \ln p(\mathbf{w}|\mathbf{t}) = \arg\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\alpha}{2\beta} \|\mathbf{w}\|_2^2$$

Regularization Framework

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{M} |\mathbf{w}_j|^q = \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_q^q$$

Regularization Framework

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|\mathbf{w}_j|^q = \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_q^q$$

- Ridge regression ($q = 2$)

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

Regularization Framework

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|\mathbf{w}_j|^q = \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_q^q$$

- Ridge regression $(q = 2)$

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- Lasso regression $(q = 1)$:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_1$$

Regularization Framework

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|\mathbf{w}_j|^q = \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_q^q$$

- Ridge regression $(q = 2)$

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- Lasso regression $(q = 1)$:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_1$$

- When $q < 1$ the problem is non-convex (more sparsity).

Solutions in simple cases

- Analysis in 1D, when $x = 1$.
  - ℓ_1: $\frac{1}{2}(w-y)^2 + \lambda|w|$,

Solutions in simple cases

- Analysis in 1D, when $x = 1$.
    - ℓ_1: $\frac{1}{2}(w - y)^2 + \lambda|w|$,
      Solution: $y \geq \lambda, w = y - \lambda$; if $y \leq -\lambda, w = y + \lambda$, else $w = 0$.

Solutions in simple cases

- Analysis in 1D, when $x = 1$.
  - ℓ_1: $\frac{1}{2}(w-y)^2 + \lambda|w|$,
    Solution: $y \geq \lambda, w = y - \lambda$; if $y \leq -\lambda, w = y + \lambda$, else $w = 0$.
  - ℓ_2: $\frac{1}{2}(w-y)^2 + \lambda w^2$,

Solutions in simple cases

- Analysis in 1D, when $x = 1$.
  - ℓ_1: $\frac{1}{2}(w - y)^2 + \lambda|w|$,
    Solution: $y \geq \lambda, w = y - \lambda$; if $y \leq -\lambda, w = y + \lambda$, else $w = 0$.
  - ℓ_2: $\frac{1}{2}(w - y)^2 + \lambda w^2$,
    Solution: $w = y/(1 + 2\lambda)$

Solutions in simple cases

- Analysis in 1D, when $x = 1$.
  - ℓ_1: $\frac{1}{2}(w - y)^2 + \lambda|w|$,
    Solution: $y \geq \lambda, w = y - \lambda$; if $y \leq -\lambda, w = y + \lambda$, else $w = 0$.
  - ℓ_2: $\frac{1}{2}(w - y)^2 + \lambda w^2$,
    Solution: $w = y/(1 + 2\lambda)$
- How about 2D case?

Notes on Regularization

- From regularized problems to constrained problems. For convex problems, a regularized problem has an equivalent constrained problem.

$$\min_x f(x) + r(x) \Leftrightarrow \min_x f(x) \text{ s.t. } x \in \mathcal{S}$$

Notes on Regularization

- From regularized problems to constrained problems. For convex problems, a regularized problem has an equivalent constrained problem.

$$\min_x f(x) + r(x) \Leftrightarrow \min_x f(x) \text{ s.t. } x \in \mathcal{S}$$

- Solving

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_q^q$$

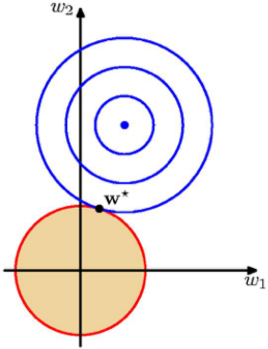
is equivalent to

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 \text{ s.t. } \|\mathbf{w}\|_q^q \leq z$$

Geometry Interpretation of Regularization

Ridge regression $(q = 2)$

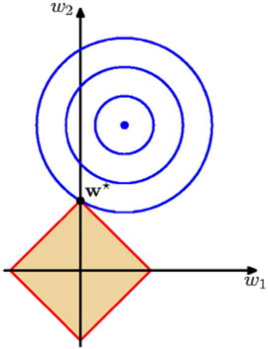
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 \text{ s.t. } \|\mathbf{w}\|_2^2 \leq z$$



Geometry Interpretation of Regularization

Lasso regression $(q = 1)$

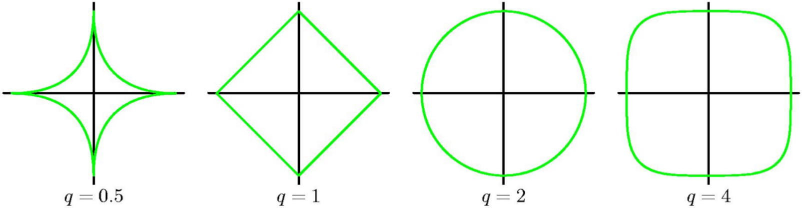
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{t}\|_2^2 \text{ s.t. } \|\mathbf{w}\|_1 \leq z$$



Geometry Interpretation of Regularization

The more general ℓ_p regularizer:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{\Phi w} - \mathbf{t}\|_2^2 \text{ s.t. } \|\mathbf{w}\|_q^q \leq z$$



Sparsity-Inducing Norms

- Lasso is commonly used for feature selection.
- Theory
  - `http://www-stat.stanford.edu/~tibs/lasso.html`
- Software
  - SLEP `http://www.public.asu.edu/~jye02/Software/SLEP`
  - FPC `http://www-stat.stanford.edu/~tibs/lasso.html`
  - L1_Ls `http://www-stat.stanford.edu/~tibs/lasso.html`

From Online Learning to Bayesian Learning

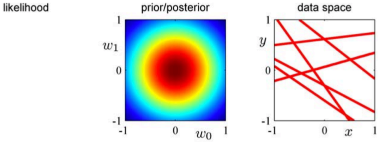
- If data points arrive sequentially, then the posterior distribution at any stage acts as the prior distribution for the subsequent data point, such that the new posterior distribution is updated sequentially.

From Online Learning to Bayesian Learning

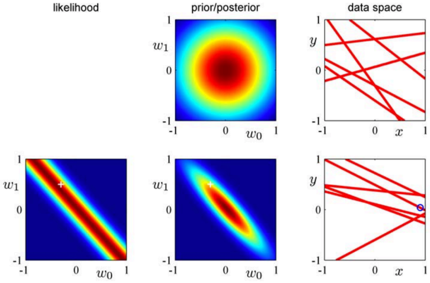
- If data points arrive sequentially, then the posterior distribution at any stage acts as the prior distribution for the subsequent data point, such that the new posterior distribution is updated sequentially.

$$
\begin{aligned}
p(\mathbf{w}|\mathcal{D};\alpha,\beta) =& p(\mathbf{w}|\mathbf{t},\mathbf{X};\alpha,\beta) \propto p(\mathbf{t}|\mathbf{X};\mathbf{w},\beta)p(\mathbf{w}|\alpha) \\
&= \prod_{i=1}^{n} p(\mathbf{t}_i|\mathbf{x}_i;\mathbf{w},\beta)p(\mathbf{w}|\alpha) \\
&= \underbrace{p(\mathbf{t}_n|\mathbf{x}_n;\mathbf{w},\beta)}_{\text{Likelihood}} \underbrace{\left[\prod_{i=1}^{n-1} p(\mathbf{t}_i|\mathbf{x}_i;\mathbf{w},\beta)p(\mathbf{w}|\alpha)\right]}_{\text{Prior}}
\end{aligned}
$$

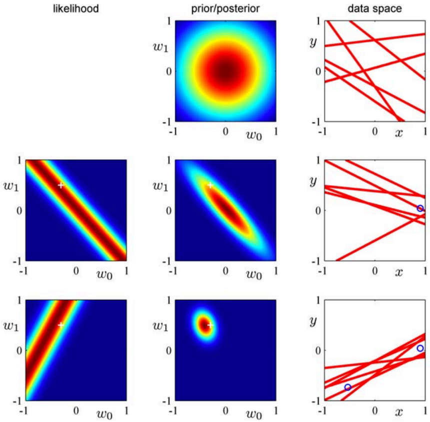
From Online Learning to Bayesian Learning



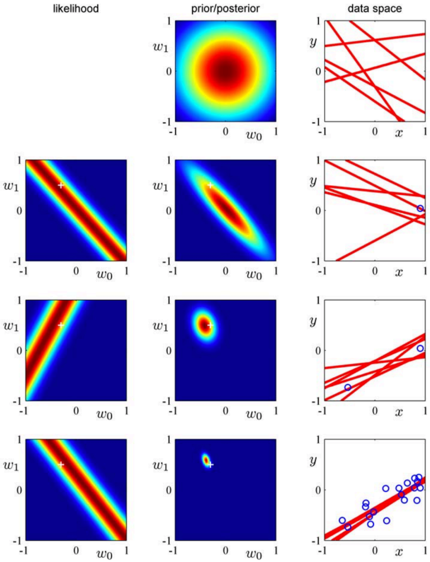
From Online Learning to Bayesian Learning



From Online Learning to Bayesian Learning



From Online Learning to Bayesian Learning



From Online Learning to Bayesian Learning

- In practice we are only interested in making predictions of t for new values of $\mathbf{x}$, and $\mathbf{w}$ is not that interesting itself.

From Online Learning to Bayesian Learning

- In practice we are only interested in making predictions of t for new values of $\mathbf{x}$, and $\mathbf{w}$ is not that interesting itself.
- Given a feature vector $\hat{\mathbf{x}}$, and training data $\mathcal{D} = \{\mathbf{X}, \mathbf{t}\}$

$$\begin{aligned} p(\hat{t}|\hat{\mathbf{x}}; \mathcal{D}, \alpha, \beta) &= \int p(\hat{t}, \mathbf{w}|\hat{\mathbf{x}}; \mathcal{D}, \alpha, \beta) d\mathbf{w} \\ &= \int p(\hat{t}|\hat{\mathbf{x}}, \mathbf{w}; \beta) p(\mathbf{w}|\mathcal{D}; \alpha, \beta) d\mathbf{w} \end{aligned}$$

where

$$p(\hat{t}|\hat{\mathbf{x}}, \mathbf{w}; \beta) = \mathcal{N}(\hat{t}|\hat{\boldsymbol{\phi}}^T \mathbf{w}, \beta^{-1}), \quad p(\mathbf{w}|\mathcal{D}; \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

From Online Learning to Bayesian Learning

- In practice we are only interested in making predictions of t for new values of $\mathbf{x}$, and $\mathbf{w}$ is not that interesting itself.
- Given a feature vector $\hat{\mathbf{x}}$, and training data $\mathcal{D} = \{\mathbf{X}, \mathbf{t}\}$

$$\begin{aligned} p(\hat{t}|\hat{\mathbf{x}}; \mathcal{D}, \alpha, \beta) &= \int p(\hat{t}, \mathbf{w}|\hat{\mathbf{x}}; \mathcal{D}, \alpha, \beta) d\mathbf{w} \\ &= \int p(\hat{t}|\hat{\mathbf{x}}, \mathbf{w}; \beta) p(\mathbf{w}|\mathcal{D}; \alpha, \beta) d\mathbf{w} \end{aligned}$$

where

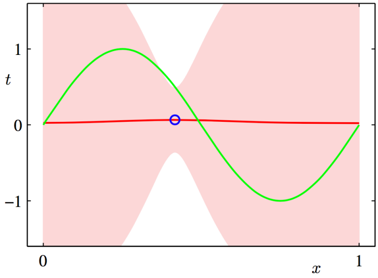
$$p(\hat{t}|\hat{\mathbf{x}}, \mathbf{w}; \beta) = \mathcal{N}(\hat{t}|\hat{\boldsymbol{\phi}}^T \mathbf{w}, \beta^{-1}), \quad p(\mathbf{w}|\mathcal{D}; \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

- You will show that (in homework):

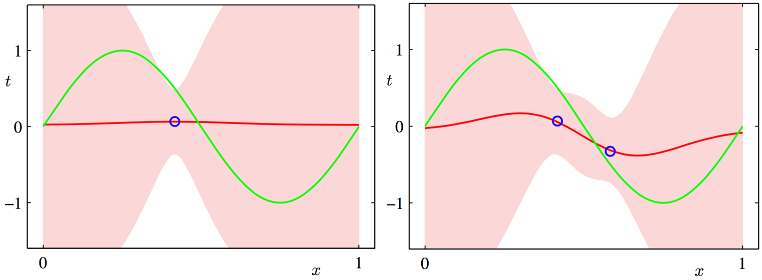
$$p(\hat{t}|\hat{\mathbf{x}}; \mathcal{D}, \alpha, \beta) = \mathcal{N}(\hat{t}|\mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

and variance $\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})$.

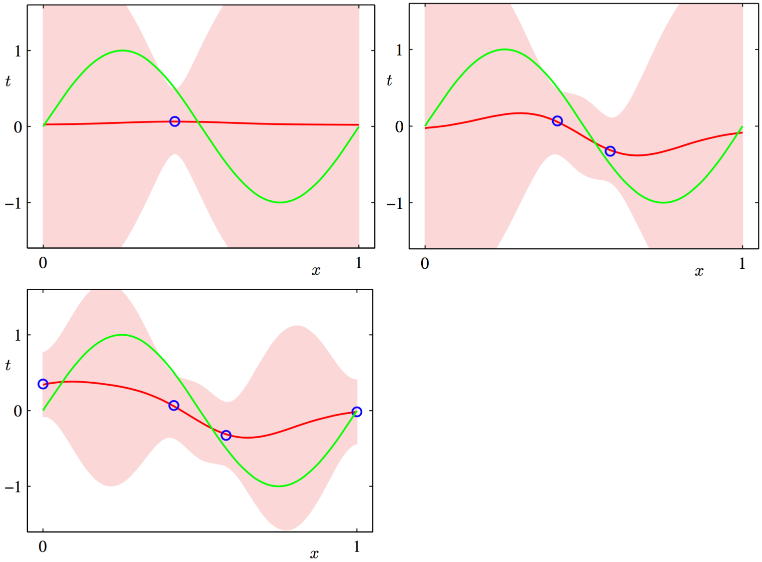
Predictive Distribution



Predictive Distribution



Predictive Distribution



Predictive Distribution