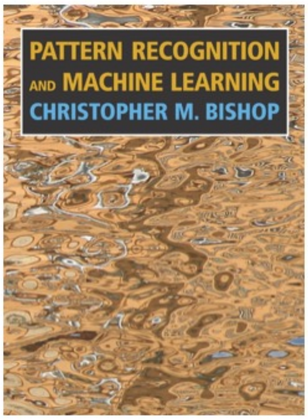
Regression: a Probabilistic View

Song Liu (song.liu@bristol.ac.uk)

Office Hour: Thursday 3-4pm

Office: GA 18

Reference



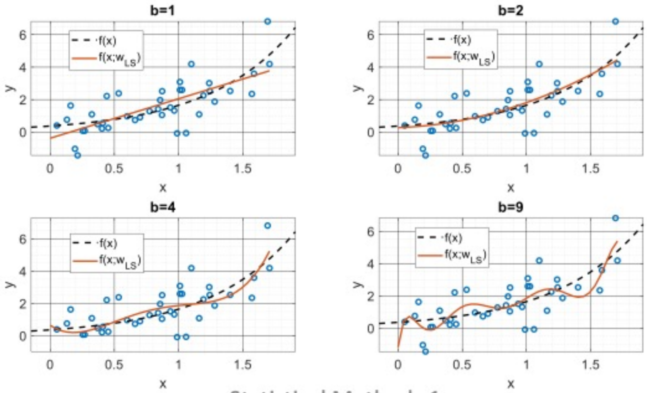
Today's class *roughly* follows Chapter 1.

Pattern Recognition and Machine Learning

Christopher Bishop, 2006

The Overfitting Issue…

- Last class, we faced a dilemma:
  - By using poly. feature, we can increase the flexibility of $f(\boldsymbol{x}; \boldsymbol{w})$.
  - The increased flexibility may also cause overfitting…



Overfitting and Regularization

- Large b causes overfitting
  - Pick a smaller b to avoid overfitting (using CV).
- What if we want to use a larger b?
  - We want the flexibility provided by high order polynomials.
- One trick we can do is called **regularization**.

$$\boldsymbol{w}_{\mathrm{LS-R}} \coloneqq \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i \in D} [y_i - f(\boldsymbol{x}_i; \boldsymbol{w})]^2 + \underbrace{\boxed{\lambda \boldsymbol{w}^\top \boldsymbol{w}}}_{\textbf{regularization term}}$$

- By adding a **regularization term** to LS Error.
- Note: $\lambda > 0$.

10/4/21 Statistical Methods 1 4

There are other cases we cannot use CV. For example, when the dataset contains non-iid samples.

Note, regularization here is introduced as a trick, but will be justified soon.

Lambda must > 0!!

Overfitting and Regularization

$$\boldsymbol{w}_{\mathrm{LS-R}} \coloneqq \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i \in D} [y_i - f(\boldsymbol{x}_i; \boldsymbol{w})]^2 + \boxed{\lambda \boldsymbol{w}^\top \boldsymbol{w}}$$

- $\boldsymbol{w}^\top \boldsymbol{w}$ is the magnitude of $\boldsymbol{w}$
- Regularization term discourages $\boldsymbol{w}$ taking large values.

- Why does the regularization help overcome overfitting?

10/4/21 Statistical Methods 1 5

Regularization simply restrict the magnitude of w by minimizing it.

In fact, any function monotone increases with the magnitude of coefficients is a reasonable choice of regularization term.

For example, $\max_i |w^{(i)}|$, i.e., the element with the maximum absolute value.

Or $\mathrm{card}(\boldsymbol{w})$, the count of non-zero elements in $\boldsymbol{w}$

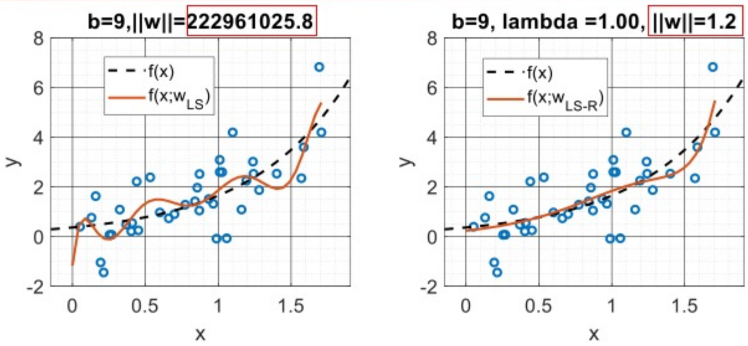
Overfitting and Regularization

- Prove that if regularization term is $\lambda \boldsymbol{w}^\top \boldsymbol{w}$,
- $\boldsymbol{w}_{\mathrm{LS-R}} := (\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{\phi}(\boldsymbol{X})^\top + \lambda \boldsymbol{I})^{-1}\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{y}^\top$,
  - $\boldsymbol{I} \in R^{b \times b}$ is identity matrix.

- $\mathrm{limit}_{\lambda \to \infty} \boldsymbol{w}_{\mathrm{LS-R}} = \boldsymbol{0}$.
  - $\mathrm{limit}_{\lambda \to \infty} f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS-R}}) = 0$.
  - As you enlarge λ, coefficients in $\boldsymbol{w}_{\mathrm{LS-R}}$ get smaller and smaller.
  - As you enlarge λ, $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS-R}})$ get flatter and flatter.
  - Which in turn reduces the complexity of $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS-R}})$

For example, ||w||_1 does not have a closed form solution for w.

Example



- $f(\boldsymbol{x}; w_{\mathrm{LS-R}})$ is much less squiggly than $f(\boldsymbol{x}; w_{\mathrm{LS}})$

  See PRML, Table 1.2 for another example

||w|| here is $\sqrt{ww^{\top}}$

The “length” of w has been shrunk by our regularizer!

Overfitting and Regularization

$$\boldsymbol{w}_{\mathrm{LS-R}} := \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i \in D} [y_i - f(\boldsymbol{x}_i; \boldsymbol{w})]^2 + \boxed{\lambda \boldsymbol{w}^\top \boldsymbol{w}}$$

- Regularization term does not have to be $\boldsymbol{w}^\top \boldsymbol{w}$
- For example, $\sum_i |w_i|$, i.e. $||\boldsymbol{w}||_1$ can be used too!

- $||\boldsymbol{w}||_1$ and $\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}$ are called “norms”.

10/4/21 Statistical Methods 1 8

Regularization simply restrict the magnitude of w by minimizing it.

In fact, any function monotone increases with the magnitude of coefficients is a reasonable choice of regularization term.

For example, $\max_i |w^{(i)}|$, i.e., the element with the maximum absolute value.

Or $\mathrm{card}(\boldsymbol{w})$, the count of non-zero elements in $\boldsymbol{w}$

Norms

- Norms are widely used in machine learning.
- a generalization of the concept "length" in Euclidean spac.
  - $\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}$ is the Euclidian distance from $\boldsymbol{w}$ to the origin.

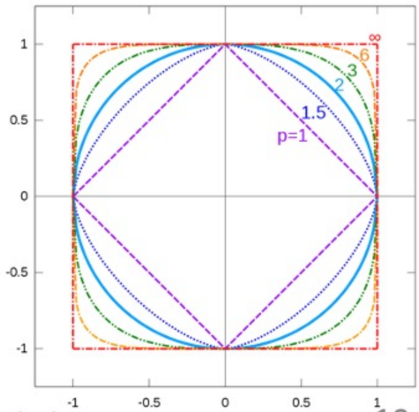
- To become a norm, a function t must satisfy
  - If $t(\boldsymbol{x}) = 0$, then $\boldsymbol{x} = \boldsymbol{0}$
  - $t(\boldsymbol{x}) + t(\boldsymbol{y}) \geq t(\boldsymbol{x} + \boldsymbol{y})$, Triangle Inequality
  - $t(a \cdot \boldsymbol{x}) = |a| \cdot t(\boldsymbol{x})$

- Matrix cookbook, page 60, 61, 62.

These three conditions also mirrors the characteristics of Euclidean distances.

L^p norms

- An important class of norms is called L^p norm.
- L^p norm for a real $p \geq 1$:
- $||\boldsymbol{x}||_p := (|x_1|^p + \cdots |x_d|^p)^{\frac{1}{p}}$
- Right: Unit "circle" defined by different L^p norms.



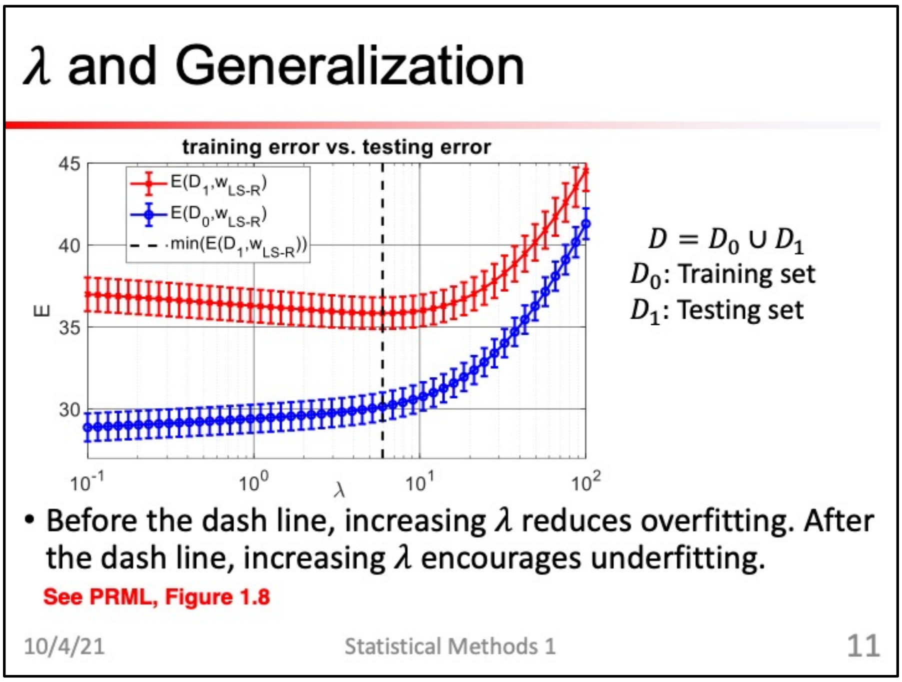
10/4/21 Statistical Methods 1 10

Euclidean norm is L2 norm, and $|\boldsymbol{w}|$ is L1 norm.

The shape of the circle tells how conservative/aggressive the distance grows in different norm.

Starting from origin, L1 distance grows the most rapidly and L infinity grows the most conservatively.

This idea is important for sparse regularization.



Basically, the regularization plot is a “reverse” of the plot over degree b we saw in the last lecture.

Problem of Regularization

- How do you choose λ?
- If we have plenty of i.i.d. data, we may choose a λ that minimizes the validation error using CV.
- However, what if we only have limited data.

- Frequentist approach does not offer a straightforward way for tuning λ. To choose λ we need to adopt **a probabilistic view of regression problem.**

Inverse Problems

- Many data science problems are inverse problems.
- We have a dataset of noisy observations D
- We want to identify some **latent**, **unobserved** data generating mechanism.

- In regression, we observe y_i which is supposedly generated by
- $y_i = g(x_i) + \epsilon$, where ϵ is some noise.
- We are interested in finding the latent function g.

The Prediction of Ceres

Inverse Problems and Posterior

- The key of solving inverse problem is to infer posterior probability distributrion $p(g|D)$.
  - The word "posterior" comes from the fact that $p(g|D)$ is a probability obtained AFTER we observe D.
  - pp. 17, PRML
- The probability of a latent, data generating mechanism, g, given our dataset D.

- Problem: How do we obtain that posterior?

Bayes' Rule (or Law, Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- You can calculate a conditional probability given its "inverse probability".
- This theorem plays a key role in Bayesian statistics.
- Let us see how it helps us to obtain posterior $p(g|D)$.

Inverting the Posterior by Bayes' Rule

- Using Bayes' Rule, we know

$$\underbrace{p(g|D)}_{\text{Posterior}} = \frac{\overbrace{p(D|g)}^{\text{Likelihood}}\overbrace{p(g)}^{\text{Prior}}}{\underbrace{p(D)}_{\text{Evidence}}}$$

- $p(g)$ is called prior: the belief of our data generation mechanism g BEFORE the observation.
- $p(D|g)$ is called likelihood as it shows how likely we observe a specific dataset D given a data generator g.

Regression using Bayes' Rule

- In regression, we want to infer $p(g|D)$, where g is the data generating function:
- $y_i = g(\boldsymbol{x}_i) + \epsilon$.
- Suppose g admits a parametric form $g(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{w})$, we only need to consider the parameter $\boldsymbol{w}$.
  - Once $\boldsymbol{w}$ is determined, f is determined.

- **Task:** Infer $p(\boldsymbol{w}|D)$
- **Bayes' Rule:** $p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{W})p(\boldsymbol{w})}{p(D)}$

Regression using Bayes' Rule

- **Task:** Infer $p(\boldsymbol{w}|D)$
- **Bayes' Rule:** $p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})p(w)}{p(D)}$
- If we assume ϵ is drawn from a Normal dist and D is IID:
- $p(D|\boldsymbol{w}) = \prod_{i \in D} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}, \sigma^2) = \prod_{i \in D} N_{y_i}(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2)$

- To compute the Bayes' rule, we also need a prior $p(\boldsymbol{w})$.
- For now, we just use a Normal dist., $p(\boldsymbol{w}) = N_{\boldsymbol{w}}(0, \sigma_{\boldsymbol{w}}^2 \mathrm{I})$.

- $\boxed{p(\boldsymbol{w}|D) = \frac{\prod_{i \in D} N_{y_i}(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2) \cdot N_{\boldsymbol{w}}(0, \sigma_{\boldsymbol{w}}^2 \mathbf{I})}{P(D)}}$

Maximum A Posteriori (MAP)

- $p(\boldsymbol{w}|D) = \boxed{\dfrac{\prod_{i \in D} N_{y_i}(f(\boldsymbol{x}_i;\boldsymbol{w}),\sigma^2) \cdot N_{\boldsymbol{w}}\left(0,\sigma_{\boldsymbol{w}}^2\mathbf{I}\right)}{P(D)}}$
- How to make a prediction?
  - Find a $\boldsymbol{w}$ that is the most likely given our dataset D!
- To get a single $\boldsymbol{w}$, we can perform a maximization of $p(\boldsymbol{w}|D)$ with respect to $\boldsymbol{w}$.
- This procedure is called Maximum A Posteriori (MAP)
- $\boldsymbol{w}_{\text{MAP}} := \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{w}|D)$

$$= \operatorname{argmax}_{\boldsymbol{w}} \prod_{i \in D} N_{y_i}(f(\boldsymbol{x}_i;\boldsymbol{w}),\sigma^2) \cdot N_{\boldsymbol{w}}(0,\sigma_{\boldsymbol{w}}^2\mathbf{I})$$

10/4/21 Statistical Methods 1 20

If you want to make a prediction, you want a w

The maximization is with respect to w. P(D) does not dependent on w, so it is ignored.

MAP is looking for the peak of your posterior distribution.

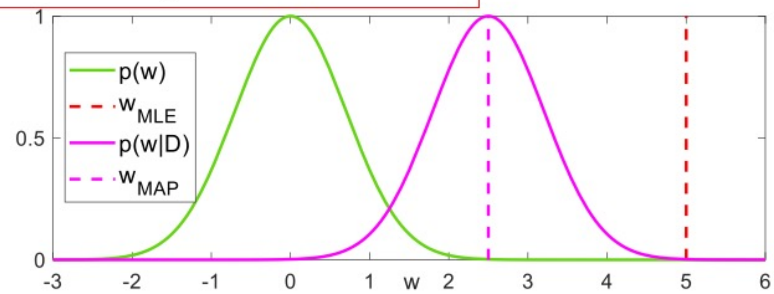
Maximum A Posteriori (MAP)

- $\boxed{\text{Prove, } \mathbf{w}_{\text{MAP}} = \boldsymbol{w}_{\text{LS-R}} \text{ using } \lambda = \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}.}$

- After getting $\boldsymbol{w}_{\text{MAP}}$, we can plug it in $f(\boldsymbol{x}; \boldsymbol{w}_{\text{MAP}})$ to make predictions.

MAP vs. MLE

- $$\mathbf{w}_{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w}|D) = \operatorname*{argmax}_{\boldsymbol{w}} p(D|\boldsymbol{w})p(\boldsymbol{w})$$

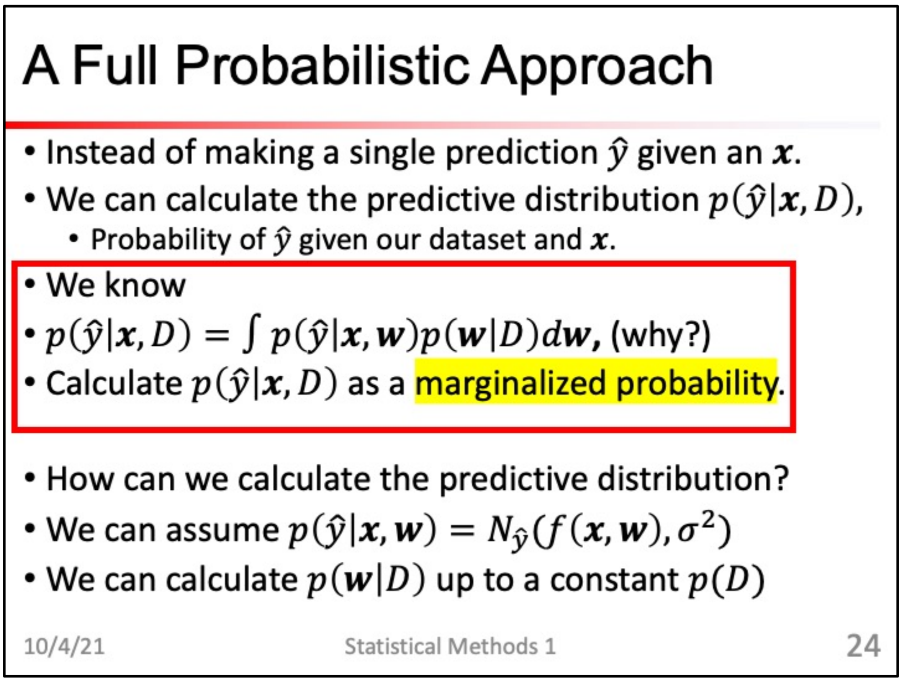


A Full Probabilistic Approach

- However, why settle with a single $\boldsymbol{w}$ when you already have access to $p(\mathbf{w}|D)$?
- Using MAP to obtain a single $\boldsymbol{w}$ for prediction **ignores the uncertainty information** represented in $p(\mathbf{w}|D)$**.**
- If not getting a single $\boldsymbol{w}$, how do we make prediction using a probability $p(\mathbf{w}|D)$?

The real argument here is, why settle with a single value when you already have a distribution of w.

It is like to evaluate the performance of this cohort, I only pick one of the student for assessment.



In full probability setting, you do not make a single prediction, you construct a **predictive distribution $\boldsymbol{p(y|x, D)}$**.

More generally, you can always think that the probabilistic model inference problem is inferring a predictive distribution $p(x|D)$.

To do so, we need can introduce a **predictive model** $p(x|D, w)$ and calculate a posterior $p(w|D)$,

Then marginalizing $p(x|D, w)p(w|D)$ with respect to w, gives us the **predictive distribution $\boldsymbol{p(y|x, D)}$**.

Calculating Predictive Distribution

likelihood prior

- $p(\boldsymbol{w}|D) \propto \prod_{i \in D} N_{y_i}\big(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2\big) \cdot N_{\boldsymbol{w}}\big(0, \sigma_{\boldsymbol{w}}^2 \mathbf{I}\big)$
- $p(\hat{y}|\boldsymbol{x}, \boldsymbol{w}) = N_{\hat{y}}(f(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$
- Suppose $f(\boldsymbol{x}; \boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$
- **Prove:**
- $\int p(\hat{y}|\boldsymbol{x}, \boldsymbol{w}) \cdot p(\boldsymbol{w}|D) d\boldsymbol{w} =$

$$N_{\hat{y}}\left[f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS-R}}), \sigma^2 + \boldsymbol{\phi}^\top(\boldsymbol{x})\sigma^2\left(\boldsymbol{\phi}\boldsymbol{\phi}^\top + \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2}\boldsymbol{I}\right)^{-1}\boldsymbol{\phi}(\boldsymbol{x})\right]$$

- Where $\boldsymbol{\phi}$ is short for $\boldsymbol{\phi}(\boldsymbol{X})$, and $\boldsymbol{w}_{\mathrm{LS-R}}$ is the LS-R solution with $\lambda = \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2}$.

The Predictive Distribution

- $p(\hat{y}|\boldsymbol{x}, D) = \int p(\hat{y}|\boldsymbol{x}, \boldsymbol{w}) \cdot p(\boldsymbol{w}|D) d\boldsymbol{w} =$
$$N_{\hat{y}}\left[f(\boldsymbol{x}; \boldsymbol{w}_{\text{LS-R}}), \sigma^2 + \boldsymbol{\phi}^\top(\boldsymbol{x})\sigma^2\left(\boldsymbol{\phi}\boldsymbol{\phi}^\top + \frac{\sigma^2}{\sigma_w^2}\boldsymbol{I}\right)^{-1}\boldsymbol{\phi}(\boldsymbol{x})\right]$$

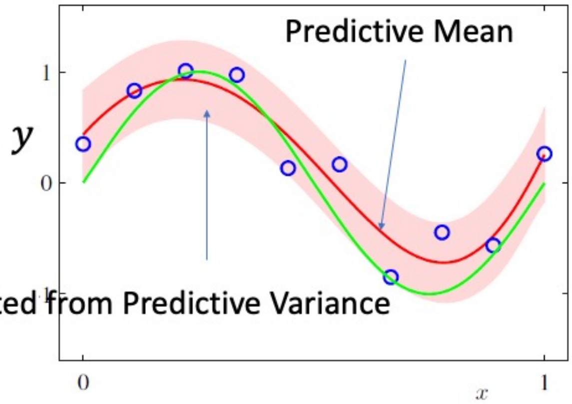
- The mean of $p(\hat{y}|\boldsymbol{x}, D)$ is the LS-R prediction!
- The idea of regularization naturally arises from both probabilistic modelling approaches.

A Full Probabilistic Approach

- With the predictive distribution $p(\hat{y}|\boldsymbol{x}, D)$, we can compute:
- Prediction: $\mathbb{E}_{p(\hat{y}|\boldsymbol{x}, D)}[\hat{y}|\boldsymbol{x}]$,
- Prediction uncertainty: $\mathrm{var}_{p(\hat{y}|\boldsymbol{x}, D)}[\hat{y}|\boldsymbol{x}]$.

- We can also use the predictive distribution to calculate other interesting expected values, as we will see later.

Example: $p(\hat{y}|x, D)$



- **PRML, Figure 1.17**

Conclusion

- We looked at "Regularized LS" from three different perspectives:
  - Regularized LS (Frequentist)
  - MAP (Semi-Bayesian)
  - Probabilistic Approach (Full Bayesian)

- However, we still have not incorporated an important concept, risk function, in our decision making process.
  - Recall, making wrong decisions has different consequences.
- Next, we talk about statistical decision making.
  - We will finally wrap up Chapter 1, PRML.

Homework

- Prove the statement on page 6
- Revisit: "The solution of $\boldsymbol{w}_{\mathrm{LS}}$ is useless if $n < d$."
  - Is this statement still true for $\boldsymbol{w}_{\mathrm{LS-R}}$?
- Prove the statement on page 21
- Prove the statement on page 25