Introduction to

Machine Learning and Data Mining

(Học máy và Khai phá dữ liệu)

Khoat Than

School of Information and Communication Technology

Hanoi University of Science and Technology

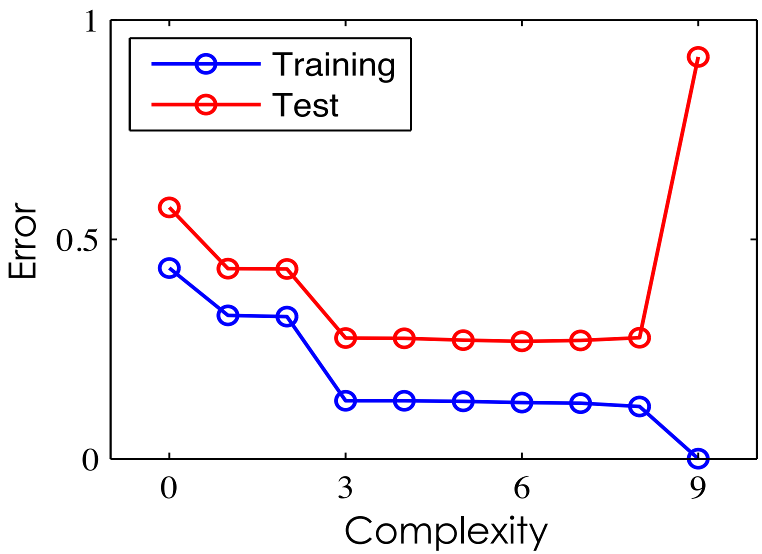
2020

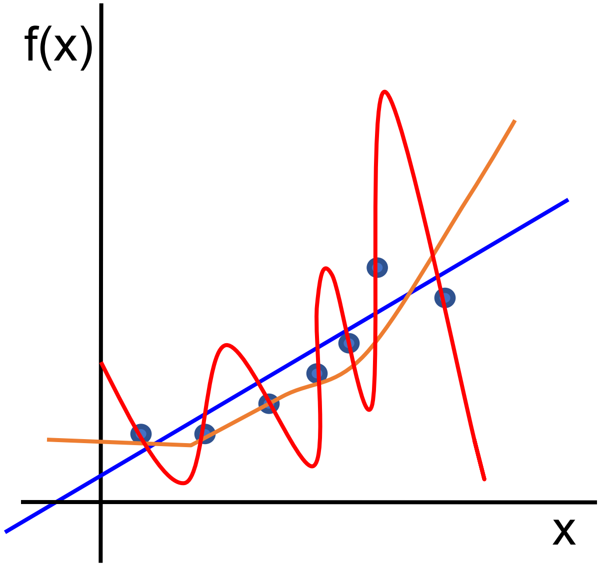
Content

- Introduction to Machine Learning & Data Mining
- Unsupervised learning
- Supervised learning
- Probabilistic modeling
- **Regularization**
- Practical advice

Revisiting overfiting

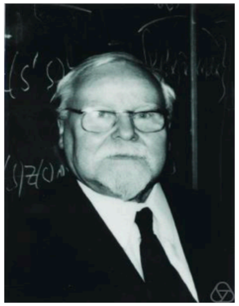
- The complexity of the learned function: y=f(x)
  - For a given training data **D**, *the more complicated f, the more possibility that f fits **D** better.*
  - For a given **D**, there exist many functions that fit **D** perfectly (i.e., no error on **D**).
  - However, those functions might generalize very badly.



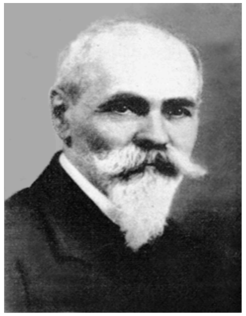


Regularization: introduction

- *Regularization* is now a popular and useful technique in ML.
- It is a technique to exploit further information to
  - Avoid overfitting in ML.
  - Solve ill-posed problems in Maths.
- The further information is often enclosed in a *penalty on the complexity* of f(x).
  - More penalty will be imposed on complex functions.
  - We prefer simpler functions among all that fit well the training data.



Tikhonov, smoothing an ill-posed problem



Zaremba, model complexity minimization

Bayes: priors over parameters

Andrew Ng: need no maths, but it prevents overfitting!

Regularization in Ridge regression

- Learning a linear regressor by ordinary least squares (OLS) from a training data **D** = {$(x_1, y_1), \ldots, (x_M, y_M)$} is reduced to the following problem:

$$w^* = \arg\min_w RSS(w, D) = \arg\min_w \sum_{(x_i, y_i) \in D} (y_i - w^T x_i)^2$$

- For Ridge regression, learning is reduced to

$$w^* = \arg\min_w RSS(w, D) + \lambda \|w\|_2^2$$

  - Where λ is a possitive constant.
  - The term $\lambda \|w\|_2^2$ plays the role as *limiting the size/complexity of w*.
  - λ allows us to trade off between fitness on **D** and generalization on future observations.
- Ridge regression is a regularized version of OLS.

Regularization: the principle

- Many ML problems are often reduced to the following optimization:

$$w^* = \text{argmin}_{w \in H} L(w, D) \qquad (1)$$

  - Where w is the parameter of the function (f) to be learned.
  - w also tell the size/complexity of that function.
  - L(w,**D**) is a *loss function* which depends on **D**. This loss shows how well function f fits **D**.

- Adding a penalty to (1), we consider

$$w^* = \text{argmin}_{w \in H} L(w, D) + \lambda . g(w) \qquad (2)$$

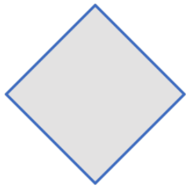
  - Where λ>0 is called *the regularization/penalty constant*.
  - g(w) measures the complexity of w.
    (it should satisfy g(w) ≥ 0)

Regularization: the principle

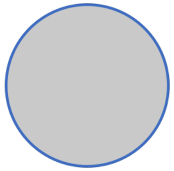
- L(w,**D**) measures the fitness of a function/model on **D**.
- The penalty (regularization) term: λ.g(w)
  - Allows to trade off the fitness on **D** and the generalization.
  - The greater λ, the heavier penalty, implying that g(w) should be small to find the best model w*.
  - In practice, λ should be neither too small nor too large.

Regularization: popular types

- G(w) often relates to some norms when w is an n-dimensional vector.
  - L_0-norm: $||w||_0$ counts the number of nonzeros in w.
  - L_1-norm: $\|w\|_1 = \sum_{i=1}^{n} |w|$



  - L_2-norm: $\|w\|_2^2 = \sum_{i=1}^{n} w_i^2$



  - L_p-norm: $\|w\|_p = \sqrt[p]{|w_1|^p + \ldots + |w_n|^p}$

Regularization in Ridge regression

- Ridge regression can be derived from OLS by adding a penalty term into the objective function when learning.
- Learning a regressor in Ridge is reduced to

$$w^* = \text{argmin}_w RSS(w, D) + \lambda \|w\|_2^2$$

  - Where λ is a possitive constant.
  - The term $\lambda \|w\|_2^2$ plays the role as regularization.
  - Large λ reduces the size of w.

Regularization in Lasso

- Lasso [Tibshirani, 1996] is a variant of OLS for linear regression by using L_1 to do regularization.
- Learning a linear regressor is reduced to

$$w^* = \operatorname{argmin}_w RSS(w, D) + \lambda \|w\|_1$$

  - Where λ is a possitive constant.
  - $\lambda \|w\|_1$ is the regularization term. Large λ reduces the size of w.
- Regularization here amounts to imposing a Laplace distribution (as prior) over each w_i, with density function:

$$p(w_i \mid \lambda) = \frac{\lambda}{2} e^{-\lambda |w_i|}$$

  - The larger λ, the more possibility that $w_i = 0$.

Regularization in SVM

- Learning a classifier in SVM is reduced to the following problem:
  - Minimize $$\frac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2}$$
  - Conditioned on $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1, \quad \forall i = 1..r$
- In the cases of noises/errors, learning is reduced to
  - Minimize $$\frac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2} + C\sum_{i=1}^{r} \xi_i$$
  - Conditioned on $$\begin{cases} y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \quad \forall i = 1..r \\ \xi_i \geq 0, \quad \forall i = 1..r \end{cases}$$
- $C(\xi_1 + \ldots + \xi_r)$ is *the regularization term*.

Regularization: MAP role

- Under some conditions, we can view regularization as

$$w^* = \operatorname{argmin}_{w \in H} \underbrace{L(w, D)}_{\text{Likelihood}} + \underbrace{\lambda . g(w)}_{\text{Prior}}$$

  - Where **D** is a sample from a probability distribution whose log likelihood is –L(w,**D**).
  - w is a random variable and follows the prior with density

$$f(w) \propto \exp\{-\lambda . g(w)\}$$

- Then $w^* = \arg\max_w \left(-L(w, D) - \lambda g(w)\right)$

$$w^* = \arg\max_w \log \Pr(D \mid w) + \log \Pr(w) = \operatorname{argmax}_w \Pr(w \mid D)$$

- As a result, regularization in fact helps us to learn an MAP solution w*.

Regularization: MAP in Ridge

- Consider the Gaussian regression model:
  - w follows a Gaussian prior: $N(w \mid 0, \sigma^2\rho^2)$.
  - Variable $f = y - w^T x$ follows the Gaussian distribution $N(f \mid 0, \rho^2, w)$ with mean 0 and variance ρ^2, and conditioned on w.
- Then the MAP estimation of f from the training data **D** is

$$w^* = \operatorname{argmax}_w \log \Pr(w \mid D) = \operatorname{argmax}_w \log\left[\Pr(D \mid w) * \Pr(w)\right]$$

$$= \operatorname{argmin}_w \sum_{(x_i, y_i)} \frac{1}{2\rho^2}\left(y_i - w^T x_i\right)^2 + \frac{1}{2\sigma^2\rho^2} w^T w - \text{constant}$$

$$= \operatorname{argmin}_w \sum_{(x_i, y_i)} \left(y_i - w^T x_i\right)^2 + \frac{1}{\sigma^2} w^T w$$

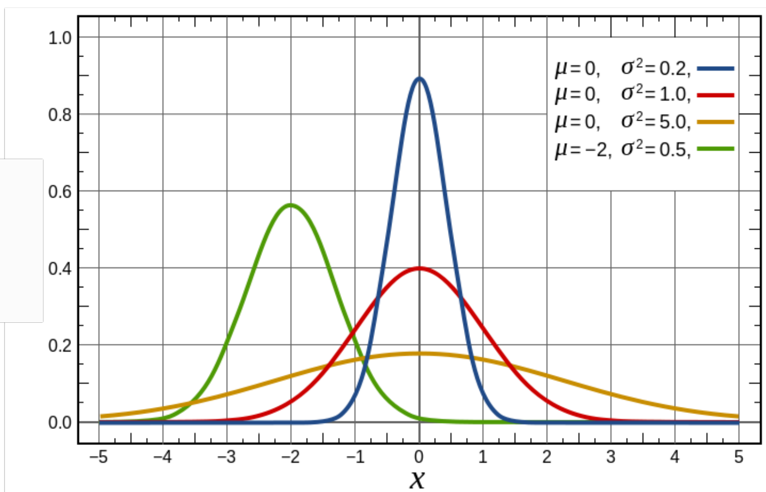
Ridge regression @@

- *Regularization using L_2 with penalty constant $\lambda = \sigma^{-2}$.*

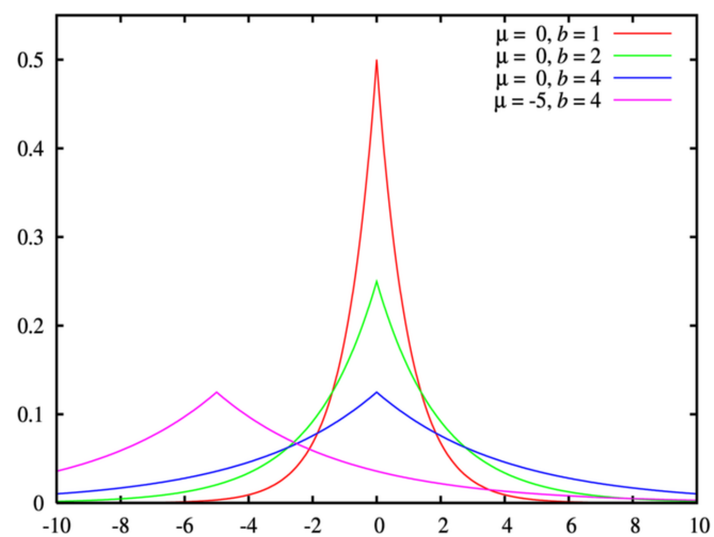
Regularization: MAP in Ridge & Lasso

- The regularization constant in Ridge: $\lambda = \sigma^{-2}$
- The regularization constant in Lasso: $\lambda = b^{-1}$
- Gaussian (left) and Laplace distribution (right)

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



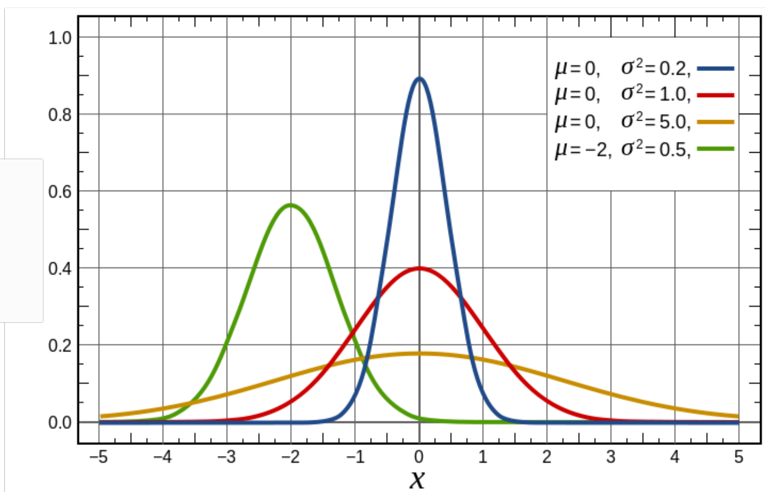
$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$



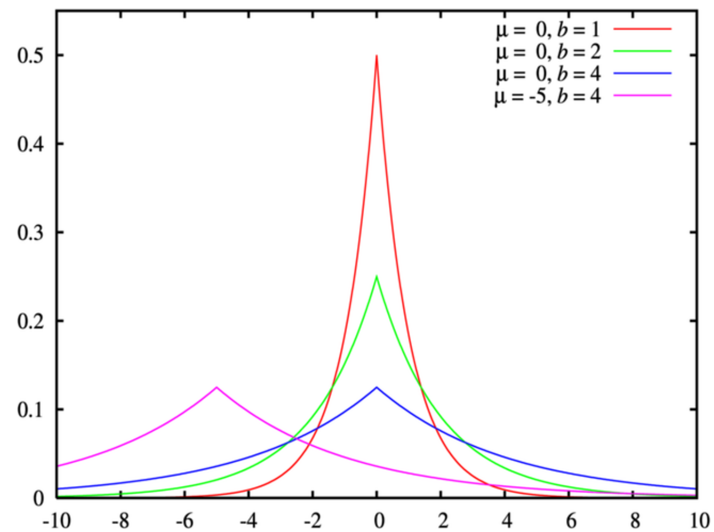
Regularization: limiting the search space

- The regularization constant in Ridge: $\lambda = \sigma^{-2}$
- The regularization constant in Lasso: $\lambda = b^{-1}$
- *The larger λ, the higher probability that x occurs around 0.*

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$f(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$



Regularization: limiting the search space

- The regularized problem:

$$w^* = \operatorname{argmin}_{w \in H} L(w, D) + \lambda . g(w) \qquad (2)$$

- A result from the optimization literature shows that (2) is equivalent to the following:

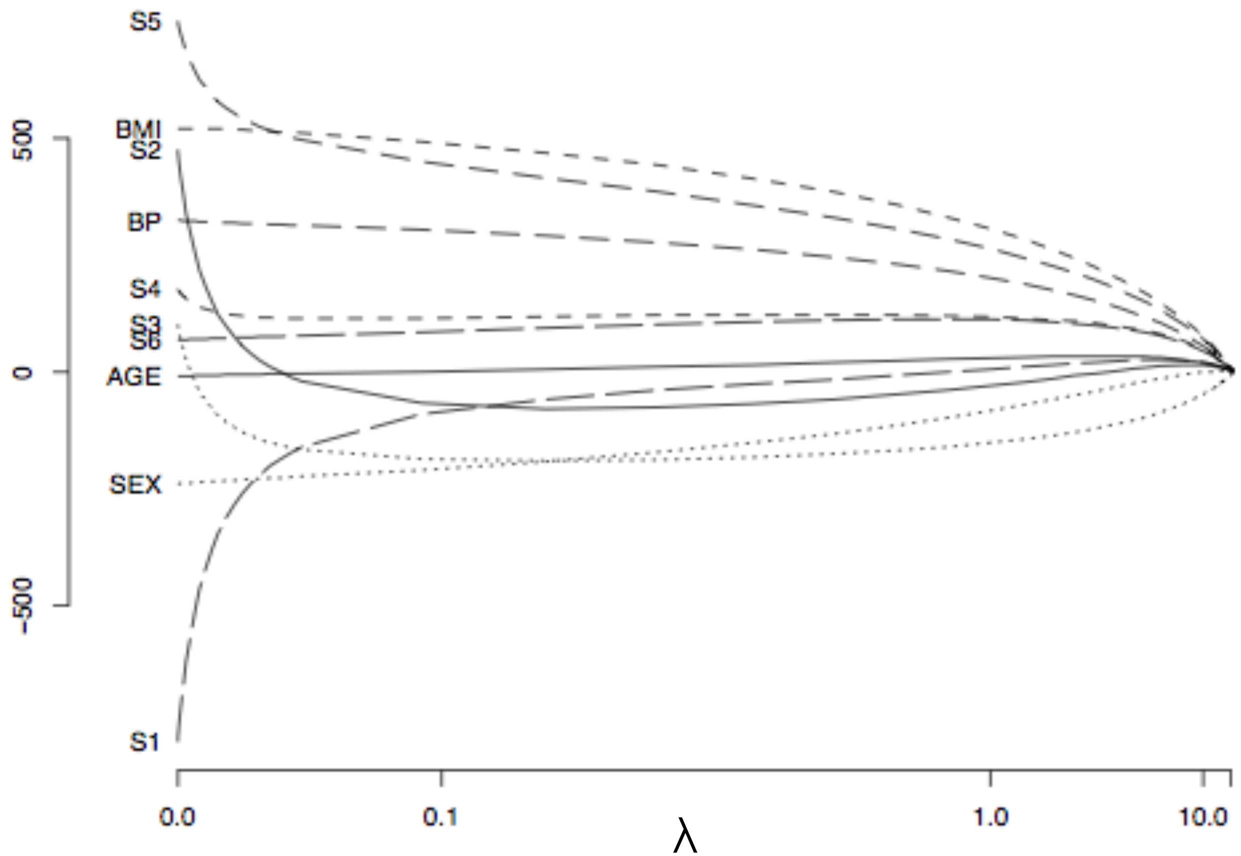
$$w^* = \operatorname{argmin}_{w \in H} L(w, D), \quad \text{s.t.} \quad g(w) \leq s \qquad (3)$$

  - For some constant s.

- *Note that the constraint of $g(w) \leq s$ plays the role as limiting the search space of w.*

Regularization: effects of λ

- Vector **w*** = (w_0, s1, s2, s3, s4, s5, s6, Age, Sex, BMI, BP) changes when λ changes in Ridge regression.
  - **w*** goes to 0 as λ increases.



Regularization: practical effectiveness

- Ridge regression was under investigation on a prostate dataset with 67 observations.
  - Performance was measured by RMSE (root mean square errors) and Correlation coefficient.

λ	0.1	1	10	100	1000	10000
RMSE	**0.74**	**0.74**	**0.74**	0.84	1.08	1.16
Correlation coeficient	0.77	0.77	**0.78**	0.76	0.74	0.73

  - Too high or too low values of λ often result in bad predictions.
  - Why??

Regularization: summary

- Advantages:
  - Avoid overfitting.
  - Limit the search space of the function to be learned.
  - Reduce bad effects from noises or errors in observations.
  - Might model data better. As an example, L_1 often work well with data/model which are inherently sparse.
- Limitations:
  - Consume time to select a good regularization constant.
  - Might pose some difficulties to design an efficient algorithm.

References

- Tibshirani, R (1996). *Regression shrinkage and selection via the Lasso*. Journal of the Royal Statistical Society, vol. 58(1), pp. 267-288.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.