

PRML Note

C01 Introduction

Zhao Yang

Department of Automation, Tsinghua University

- Generalization is a central goal in PR
- The Original input variables are typically preprocessed. The Test data must be preprocessed using the same steps as training data.
- feature extraction $\left\{ \begin{array}{l} \text{PR problem easy to solve} \\ \text{speed up computation} \end{array} \right.$
- supervised learning $\left\{ \begin{array}{l} \text{classification} \\ \text{regression} \end{array} \right.$
- unsupervised learning $\left\{ \begin{array}{l} \text{clustering} \\ \text{density estimation} \\ \text{dimensions reduction} \end{array} \right.$

1 Example: Polynomial Curve Fitting

- Linear Model:

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M = \sum_{j=0}^M w_jx^j \quad (1)$$

\mathbf{w} determined by

$$\operatorname{argmin}_{\mathbf{w}^*} E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

which is called error function

M : model comparison or model selection

$$E_{RMS} = \sqrt{\frac{2E(\mathbf{w}^*)}{N}}$$

in which the division by N allows us to compare different size of data sets.

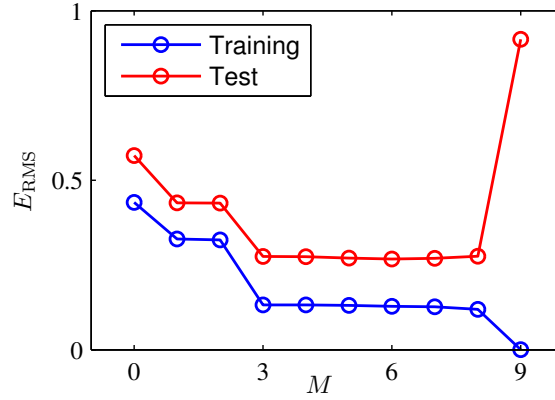


Figure 1: Graphs of the RMS error evaluated on the training set and on an independent test set for various values of M

- Large value of M \rightarrow Polynomials flexible \rightarrow increasing tuned to the random noises on target values
- The larger the data set, the more complex the model that we can afford to fit
- By adopting a Bayesian approach, over-fitting can be avoided.
In a Bayesian model, the effective number of parameters adapts automatically to the size of the data set
- regularization:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, (\mathbf{w})) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

1. shrinkage in statistics
 2. ridge regression
 3. weight decay in NN
- $\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$
 1. w_0^2 is often omitted. if not, the result will depend on the choice of the origin for the target variable
 2. $\frac{\lambda_1}{2} w_0^2 + \frac{\lambda_2}{2} (w_1^2 + \dots + w_M^2)$ is also OK
 - data set $\left\{ \begin{array}{l} \text{training set} \rightarrow \mathbf{w} \\ \text{validation set (hold-out set)} \\ \rightarrow M, \lambda \text{ which is too wasteful of data set} \\ \text{test set} \end{array} \right.$

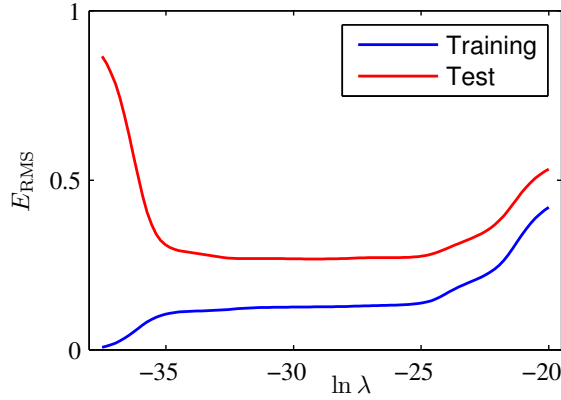


Figure 2: Graphs of the RMS error evaluated on the training set and on an independent test set for $\ln \lambda$

2 Probability Theory

- the rules of Probability

1. sum rule: $P(x) = \sum_Y P(x, Y)$
2. product rule: $P(x, Y) = P(Y|x)P(x)$
in which $P(x)$ is called marginal probability and $P(Y|x)$ is called conditional probability.

- Bayes' Theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_Y P(X|Y)P(Y)}$$

- a nonlinear change of variable $x = g(y)$
consider a probability density $p_x(x)$ that correspond to a density $p_y(y)$.
Observations falling in the range $(x, x + \delta x)$ will be transformed into the range $(y, y + \delta y)$ where $p_x(x) \simeq p_y(y)$

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right| = p_x(g(y)) |g'(y)|$$

So, the concept of the max of a probability density is dependent on the choice of variable.

- Expectations: $\mathbb{E}[f] = \sum_x p(x)f(x) \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$
Conditional Expectations: $\mathbb{E}[f|y] = \sum_x p(x|y)f(x)$

- Variances:

$$Var[f] = \mathbb{E}\{(f(x) - \mathbb{E}[f(x)])^2\} = \mathbb{E}\{f^2(x)\} - \mathbb{E}^2\{f(x)\}$$

Corvariances:

$$cov(x, y) = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] = \mathbb{E}_{xy}\{xy\} - \mathbb{E}\{x\}\mathbb{E}\{y\}$$

- a prior probability distribution $p(\mathbf{w})$
Observed data $D = \{t_1, t_2, \dots, t_N\}$ is expressed through the conditional probability $p(D|\mathbf{w})$

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})P(\mathbf{w})}{p(D)}$$

in which $p(D|\mathbf{w})$ is called likelihood function which expresses how probable the observed data set is for different settings of the parameters \mathbf{w} and $p(D)$ is called normalization constant which is $p(D) = \int p(D|\mathbf{w})p(\mathbf{w})d\mathbf{w}$

- A data set of observations $X = (x_1, x_2, \dots, x_N)^T$, i.i.d, the likelihood function is

$$p(X|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

and the log likelihood function can be written in the form

$$\ln p(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln(\sigma^2) - \frac{N}{2} \ln(2\pi)$$

- find the μ and σ^2 to max this likelihood function and the maximum likelihood solution is

$$\begin{cases} \mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n \\ \sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2 \end{cases}$$

and the expectation of it is

$$\begin{cases} \mathbb{E}\{\mu_{ML}\} = \mu \\ \mathbb{E}\{\sigma_{ML}^2\} = (\frac{N-1}{N})\sigma^2 \end{cases}$$

so

$$\tilde{\sigma}^2 = \frac{N}{N-1} \sigma_{ML}^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{ML})^2$$

- bootstrap: suppose data set consists of N data points $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, we can drawing N data points at random from \mathbf{X} to create a new data set \mathbf{X}_B . This process can be repeated L times
- maximum posterior(MAP): suppose training data comprising N input values $\mathbf{x} = (x_1, x_2, \dots, x_N)^T$ and their corresponding target values $\mathbf{t} = (t_1, t_2, \dots, t_N)^T$. We have

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \quad (2)$$

where $y(x, \mathbf{w})$ is given by equation (1). So the likelihood function is given by

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

and the log likelihood function is

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi$$

If we consider a prior distribution over \mathbf{w} ,

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1} \mathbf{I})$$

then the posterior distribution for \mathbf{w} is

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

and we should maximize the log posterior function

$$\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$

- Bayesian curve fitting: Given training data \mathbf{x} and \mathbf{t} , along with a new point x , we want to predict the value of t ,

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}$$

3 Model Selection

- data is plentiful: train a range of models or a model with a range of values for its complexity params.
- data is limited: cross-validation
- An ideal approach:
 1. only rely on the training data
 2. allow multiple hyperparameters and model types to be compared in a single training run

therefore, find a measure of performance which depends only on the training data, not suffer from bias to over-fitting.

4 The Curse of Dimensionality

- Not all intuitions developed in spaces of low dimensionality will generalize to spaces of many dimensions
- real data
 1. often be confined in lower effective dimensionality
 2. exhibit smooth properties

5 Decision Theory

- For 2-classes, minimize

$$\begin{aligned} p(\text{mistake}) &= p(x \in R_1, C_2) + p(x \in R_2, C_1) \\ &= \int_{R_1} p(x, C_2) dx + \int_{R_2} p(x, C_1) dx \end{aligned}$$

if $p(x, C_1) > p(x, C_2)$, which is the same as $p(C_1|x)p(x) > p(C_2|x)p(x)$ i.e. $p(C_1|x) > p(C_2|x)$, $x \rightarrow C_1$

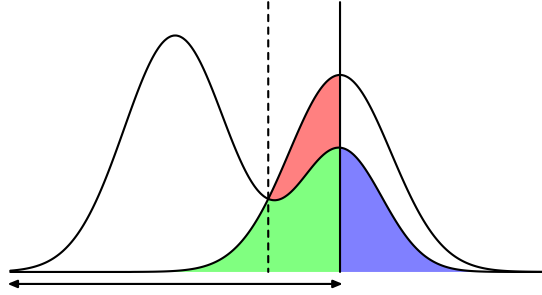


Figure 3: The joint probability of two classes

- For k -classes, maximize

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^K \int_{R_k} p(x, C_k) dx \\ &= \sum_{k=1}^K \int_{R_k} p(C_k|x)p(x) dx \end{aligned}$$

find the max $p(C_k|x)$ and $x \rightarrow C_k$

- reject option: if $p(C_k|x) < \theta$, do not classify x
 1. $\theta = 1$, reject all samples
 2. For k -classes, $\theta < \frac{1}{k}$, No sample is rejected
- Generative model: find the joint distribution $p(C_k, x)$ and then find the posterior distribution $p(C_k|x)$ to make decision using decision theory. This can be used to detect new data point with low probability, which is known as outlier detection or novelty detection.
- Discriminative model: find the posterior distribution directly, and use it to make decision.
- Discriminant function: only a function, the input is x and the output is the class label.
- Compensating for class prior: In training data $n(C_1) = a$ and $n(C_2) = b$, if $a \gg b$, it is not likely to generalize well. From the training data we can get a balanced data set. From the balanced data we can get the posterior $p(C_1|x)$ and $p(C_2|x)$. And the goal data set we want to apply to is $n(C_1) = a'$ and $n(C_2) = b'$, so the posterior we use is

$$\tilde{p}(C_1|x) = \frac{p(C_1|x)}{\frac{a}{a+b}} \frac{a'}{a' + b'}$$

and

$$\tilde{p}(C_2|x) = \frac{p(C_2|x)}{\frac{b}{a+b}} \frac{b'}{a' + b'}$$

then normalize to ensure

$$\tilde{p}(C_1|x) + \tilde{p}(C_2|x) = 1$$

Explanation:

$$\frac{p(C_1|x)}{\frac{a}{a+b}} \propto \text{likelihood function} = p(x|C_1)$$

- Combining models: $p(\mathbf{x}_I, \mathbf{x}_B|C_k) = p(\mathbf{x}_I|C_k)p(\mathbf{x}_B|C_k)$, so we can get

$$\begin{aligned} p(C_k|\mathbf{x}_I, \mathbf{x}_B) &\propto p(\mathbf{x}_I, \mathbf{x}_B|C_k)p(C_k) \\ &\propto p(\mathbf{x}_I|C_k)p(\mathbf{x}_B|C_k)p(C_k) \\ &\propto \frac{p(C_k|\mathbf{x}_I)p(C_k|\mathbf{x}_B)}{p(C_k)} \end{aligned}$$

- Calculus of Variations: if $J(y) = \int_{x_1}^{x_2} \mathcal{L}(x, y(x), y'(x))dx$, we have

$$\frac{\partial \mathcal{L}}{\partial f} - \frac{d}{dx} \frac{\partial \mathcal{L}}{\partial f'} = 0$$

which is called Euler-Lagrange Equation.

- Incurring a Loss function $\mathcal{L}(t, y(\mathbf{x}))$, the average, or expected, loss is given by

$$\mathbb{E}[L] = \iint \mathcal{L}(t, y(\mathbf{x})) p(\mathbf{x}, t) d\mathbf{x} dt \quad (3)$$

A common choice of loss function is $\mathcal{L}(t, y(\mathbf{x})) = \{y(\mathbf{x}) - t\}^2$, using the Calculus of Variations

$$\frac{\partial \mathbb{E}[L]}{\partial y(\mathbf{x})} = 2 \int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t) dt = 0$$

So we have

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) dt}{p(\mathbf{x})} \int t p(t|\mathbf{x}) dt = \mathbb{E}_t[t|\mathbf{x}]$$

- from equation (3), we can rewrite to

$$\begin{aligned} \mathbb{E}[L] &= \iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt \\ &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + \int \text{Var}[t|\mathbf{x}] p(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (4)$$

The process to get the equation (4) is below:

1. $\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2$
2. $p(\mathbf{x}, t) = p(t|\mathbf{x})p(\mathbf{x})$
- 3.

$$\begin{aligned} \mathbb{E}[L]_1 &= \iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}, t) d\mathbf{x} dt \\ &= \iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \\ &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}) \left(\int p(t|\mathbf{x}) dt \right) d\mathbf{x} \\ &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

- 4.

$$\begin{aligned} \mathbb{E}[L]_2 &= \iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\} \{\mathbb{E}[t|\mathbf{x}] - t\} p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \\ &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\} p(\mathbf{x}) \left(\int \{\mathbb{E}[t|\mathbf{x}] - t\} p(t|\mathbf{x}) dt \right) d\mathbf{x} \\ &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\} p(\mathbf{x}) (\mathbb{E}[t|\mathbf{x}] - \int t p(t|\mathbf{x}) dt) d\mathbf{x} \\ &= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\} p(\mathbf{x}) (\mathbb{E}[t|\mathbf{x}] - \mathbb{E}[t|\mathbf{x}]) d\mathbf{x} \\ &= 0 \end{aligned}$$

5.

$$\begin{aligned}
\mathbb{E}[L]_3 &= \iint \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \\
&= \iint \{\mathbb{E}^2[t|\mathbf{x}] - 2t\mathbb{E}[t|\mathbf{x}] + t^2\} p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \\
&= \int \{-\mathbb{E}^2[t|\mathbf{x}] + \mathbb{E}[t^2|\mathbf{x}]\} p(\mathbf{x}) d\mathbf{x} \\
&= \int \text{Var}[t|\mathbf{x}] p(\mathbf{x}) d\mathbf{x}
\end{aligned}$$

$$6. \mathbb{E}[L] = \mathbb{E}[L]_1 + \mathbb{E}[L]_2 + \mathbb{E}[L]_3$$

- Squared loss can lead to very poor results.
- The Minkowski loss $|y(\mathbf{x}) - t|^q$

6 Information Theory

- $h(x) = -\log_2 p(x)$
- the entropy of the random variable $x(\text{bit})$: $H[x] = -\sum_x p(x) \log_2 p(x)$ and the entropy of the random variable $x(\text{nat})$: $H[x] = -\frac{1}{\ln 2} \sum_x p(x) \ln p(x)$
- if $p(x) = 0$, we have $p(x) \log_2 p(x) = 0$
- The conditional entropy of y given x :

$$H[y|x] = - \iint p(y, x) \ln p(y|x) dy dx$$

$$\text{and } H[x, y] = H[y|x] + H[x]$$

- relative entropy or KL divergence between the distribution $p(x)$ and $q(x)$

$$\begin{aligned}
KL(p||q) &= - \int p(x) \ln q(x) dx - \left(- \int p(x) \ln p(x) dx \right) \\
&= - \int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx
\end{aligned} \tag{5}$$

Note that $KL(p||q) \neq KL(q||p)$ and $KL(p||q) \geq 0$ with equality, if and only if $p(x) = q(x)$

- the mutual information between the variables x and y

$$\begin{aligned}
I[x, y] &= KL(p(x, y) || p(x)p(y)) \\
&= - \iint p(x, y) \ln \left\{ \frac{p(x)p(y)}{p(x, y)} \right\} dx dy
\end{aligned} \tag{6}$$

Note that $I[x, y] \geq 0$ with equality, if and only if x and y are independent and we have $I[x, y] = H[x] - H[x|y] = [y] - H[y|x]$