Notes 1-Introduction

Junda Du

2019-8-5

1 Example: Polynomial Curve Fitting

1. Task: Find a linear model $y(x, \mathbf{w}) = w_0 + w_1 \cdots + w_M$, fitting the training set to give small values for the test error.

2. Definition:

   - Linear model is linear in the unknown parameters.
   - Sum of squares of the errors is $E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$
   - RMS(root-mean-square) error is $E_{RMS} = \sqrt{2E(\mathbf{w})/N}$, which allows us to compare different size of data sets on an equal footing, and ensures that E_{RMS} is measureed on the same scale as the target variable t.

3. Train: Select different M for $N = 15$ data points.

   - Values of M in the range $0 \leq M \leq 2$ give relatively large values of test error, which attributed to the fact that the corresponding polynomials are rather inflexible and incapable of fitting the dataset.
   - Values of M in the range $3 \leq M \leq 8$ give small values for the test error.
   - For $M = 9$, the training set error goes to zero, but the test error has become very large. That is known as over-fitting.Essentially, the coefficients become finely tuned to the train data by developing large positive and negative values, so it matches each data points exactly even for the noise.

4. Solution for over-fitting

   - Increasing the size of the data set if the data size is flexiable.
   - Regularization, which involves adding a penalty term to the error function to discourgae the coefficients from reaching large values.? $\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$, where $\|\mathbf{w}\| = \mathbf{w}^T\mathbf{w}$ and

the w_0^2 is often omitted. The $ln\lambda$ is often tuned linearly.?quadratic regularizer is calledridge regression

- ?Bayesian approach to avoid the over-fitting problem even if the number of parameters greatly exceeds the number of data points. In a Bayesian model the effective number of parameters adapt automatically to the size of data set.
- ? chapter 3 – the number of parameters is notnecessarily the most appropriate measure of model complexity.

2 Probablity Theory

Identities:

- Sum rule: $p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$, that is $p(X) = \sum_{Y} p(X, Y)$ or $p(x) = \int p(x, y)dy$
- Product rule: $p(X = x_i, Y = y_j) = p(Y = y_j|X = x_i)p(X = x_i)$, that is $p(X, Y) = p(Y|X)p(X)$
- *Baysian' theorem*: $p(X) = \sum_{Y} p(X|Y)p(Y)$
- $p(X) = \sum_{Y} p(X|Y)p(Y)$
- $p(X, Y) = p(X)p(Y) \Leftrightarrow X, Y$ are independent $\Leftrightarrow p(Y|X) = p(Y)$

2.1 Probablity densities

Points:

- Definition of probability density $p(x)$: If the probability of a real-valued varibale x falling in the interval$(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$,then $p(x)$ is called the *probablity density* over x
- Two conditions $p(x)$ must satisfy:$p(x) \geq 0, \int_{-\infty}^{\infty} p(x)dx = 1$, which holds for univariable and multivariable
- For nonlinear change of variable $x = g(y), p(y) = p(x)\left|\frac{dx}{dy}\right| = p(g(y))|g'(y)|$.Because $p(x)\delta x = p(y)\delta y$.
- Definition of cumulative distribution function $P(z) = \int_{-\infty}^{z} p(x)dx$ and $P'(z) = p(x)$

2.2 Expectations and covariances

Identities:

- Expectation of $f(x)$: $E[f] = \sum_x p(x)f(x)$ for a discrete distribution $E[f] = \int p(x)f(x)dx$ for continuous varibales.
- $E_x[f(x,y)] = \sum\limits_x f(x,y)p(x,y)$
- $E_x[f|y] = \sum\limits_x p(x|y)f(x)$
- Variance of $f(x)$: $var[f] = E[(f(x) - E[f(x)])^2] = E[f(x)^2] - E[f(x)]^2$
- Covariance of $\mathbf{x}$ and $\mathbf{y}$ $cov[\mathbf{x},\mathbf{y}] = E_{x,y}[\{\mathbf{x} - E[\mathbf{x}]\}\{\mathbf{y}^T - E[\mathbf{y}^T]\}] = E_{x,y}[\mathbf{x}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{y}^T]$, and ignore T if x,y are univaribale.
- $E[X] = \sum\limits_{i=1}^{N} c_i E[x_i]$, for $X = \sum\limits_{i=1}^{N} c_i x_i$ and c_i is constant
  $= \prod\limits_{i=1}^{N} c_i E[x_i]$, for $X = \prod\limits_{i=1}^{N} c_i x_i, c_i$ is constant and $x_1, \ldots, x_N$ are independent.
- $var[X] = \sum\limits_{i=1}^{N} c_i^2 var[x_i]$, for $X = \sum\limits_{i=1}^{N} c_i x_i$, c_i is constant and $x_1, \ldots, x_N$ are independent.

2.3 Bayesian probablities

$p(\mathbf{w}|D) = \dfrac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)}$, where D is the observed data $\{t_1, \ldots, t_N\}$ and $\mathbf{w}$ is the parameters of model.
Explanations:

- $p(\mathbf{w})$ is the prior probablity, $p(D|\mathbf{w})$ is the conditional probablity or the likelihood function, $p(\mathbf{w}|D)$ is the posterior probablity.
- posterior $\propto$ likelihood $\times$ prior
- Denominatorin Bayes' theorem $p(D) = \int p(D|\mathbf{w})p(\mathbf{w})d\mathbf{w}$
- $p(D|\mathbf{w})$ can be understood as the probablity of observed data appearing when the model parameters are fixed.

2.4 The Gaussian distribution

Points:

- Definition of Gaussion distribution:
  $N(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$ for univariable,
  $N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$ for multivaribale.

- Attribute: $E[x] = \int_{-\infty}^{\infty} N(x|\mu,\sigma^2)xdx = \mu$,$E[x^2] = \int_{-\infty}^{\infty} N(x|\mu,\sigma^2)x^2dx = \mu^2+\sigma^2$,$var(x) = E[x^2] - E[x]^2 = \sigma^2$

Estimate and evaluate($\mathbf{x} = (x_1,\ldots,x_n)^T$ are independent and identically distributed)

1. Use max likelihood to estimate μ and σ according to observed data: $p(\mathbf{x}|\mu\sigma^2) = \prod_{n=1}^{N} N(x_n|\mu,\sigma^2)$.Differentiate $ln(p)$ with respect to μ and $\sigma^{\mathbf{2}}$, then get $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and $\sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$

2. $E[\mu_{ML}] = \frac{1}{N}\sum_{n=1}^{N} E[x_n] = \mu$,

$$
\begin{aligned}
E[\sigma^2_{ML}] &= E[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2] = \frac{1}{N}\sum_{n=1}^{N}(E[(x_n-\mu_{ML})^2]) \\
&= \frac{1}{N}\sum_{n=1}^{N}(var[x_n-\mu_{ML}] + E[x_n-\mu_{ML}]^2) \\
&= \frac{1}{N}\sum_{n=1}^{N}(var[\frac{N-1}{N}x_n \text{-} \frac{1}{N}(\sum_{i=1,i\neq n}^{N} x_i)] + E[x_n-\mu_{ML}]^2) \\
&\overset{i.i.d.}{=\!=} \frac{1}{N}\sum_{n=1}^{N}(\frac{(N-1)^2}{N^2}var[x_n] + \frac{1}{N^2}(\sum_{i=1,i\neq n}^{N} var[x_i])) + (E[x_n]-E[\mu_{ML}])^2 \\
&= \frac{(N-1)^2+(N-1)}{N^2}\sigma^2 = \frac{(N-1)}{N}\sigma^2
\end{aligned}
$$

2.5 Curve fitting re-visited(Estimate the distribution of model parameters)

Assuming that $t_n \sim N(t_n|y(x_n,\mathbf{w}),\beta^{-1}), \mathbf{w} \sim N(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})$
According to Baysian's probablity: posterior probablity $p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$
Transform equivalently:

$$
\begin{aligned}
MAP &\Leftrightarrow maxln(p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta)) \\
&\Leftrightarrow max -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \\
&\Leftrightarrow min\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \\
&\Leftrightarrow min\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}
\end{aligned}
$$

2.6 Bayesian curve fitting(Estimate the distribution of prediction given the input)

- Normalize the $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$ in last section, geting the distribution of model parameters.
- $p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})d\mathbf{w}$
- The result is 1.69-1.72 on the page 31 of PRML.

3 Model Selection

If there are many parameters governing the model complexity, it's unreasonable to use cross-validation because exploring combinations of settings for such parameters require lots of training runs.So we need a better approach which allow multiple hyperparameters and model types to be compared in a single training run.We therefore need to find a measure which does not suffer from bias due to over-fitting.
Historically various 'information criteria' have been proposed that attempt to correct for the bias of maximum likelihood by the addition of a penalty term tocompensate for the over-fitting of more complex models.
One example is maximizing $lnp(D|\mathbf{w}_{ML}) - M$, where M is the number of adjustable parameters in the model.

4 The curse of Dimensionality

What is the relationship between the curse and the concentrated density of sphere with large dimension ?

5 Decision Theory

Decision theory is based on Baysian's therom, and it's a new method of understanding the Baysian's therom.[Different from the estimating the model parameters.]
This time we use an example that given the result of examination, our goal is to decide which of the two classed to assign to the image.The therom is also

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})}$$

5.1 Minimizing the misclassification

$p(mistake) = p(\boldsymbol{x} \in R_1, C_2) + p(\boldsymbol{x} \in R_2, C_1) = \int_{R_1} p(\boldsymbol{x}, C_2)d\boldsymbol{x} + \int_{R_2} p(\boldsymbol{x}, C_1)d\boldsymbol{x}$, where $p(\boldsymbol{x}, C_k) = p(C_k|\boldsymbol{x})p(\boldsymbol{x})$

For the more general case of K classes, it is slightly easier to maximize the probability of being correct, which is given by

$$p(correct) = \sum_{k=1}^{K} p(\boldsymbol{x} \in R_k, C_k) = \sum_{k=1}^{K} \int_{R_k} p(\boldsymbol{x}, C_k) d\boldsymbol{x}.$$

5.2 Minimizing the expected loss

Assign each $\boldsymbol{x}$ to the class j for which the quantity $\sum_k L_{kj} p(C_K|\boldsymbol{x})$ is a minimum.

5.3 The reject option

Inputs x such that the larger of the two posterior probabilities is less than or equal to some threshold θ will be rejected.

5.4 Inference and decision

why (a) and (b) are different? First get $p(C_k|\boldsymbol{x})$,then decide.

5.5 Loss functions for regression

- Given the input $\boldsymbol{x}$ and target t, the expected loss is $E[L] = \iint L(t, y(\boldsymbol{x})) p(\boldsymbol{x}, t) d\boldsymbol{x} dt$
- $L(t, y(\boldsymbol{x})) = |y(\boldsymbol{x}) - t|^q$

Our goal is to choose $y(\boldsymbol{x})$ to minimize $E[L]$, example of $q = 2$:

$$\begin{aligned} E[L] &= \iint (y(\boldsymbol{x}) - t)^2 p(\boldsymbol{x}, t) d\boldsymbol{x} dt \\ \frac{\delta E[L]}{\delta y(\boldsymbol{x})} &= 2 \int \{y(\boldsymbol{x}) - t\} p(\boldsymbol{x}, t) dt = 0 \\ y(\boldsymbol{x}) &= \int t p(t|\boldsymbol{x}) dt = E_t[t|\boldsymbol{x}] \end{aligned}$$

6 Information Theory

Points

- Information and Entroy
  1. Define the information of x: $h(x) = -log_2 p(x)$ to satisfy the conditions $h(x) \geq 0$ and $h(x, y) = h(x) + h(y)$, where x, y are unrelated.
  2. The average amount of information being transmitted is called the entroy of variable x, $H[x] = -\sum_x p(x) log_2 p(x)$.
- Maximizing entroy

1. For discrete variable X, $H[p] = -\sum_i p(x_i) ln\, p(x_i), \sum_i p(x_i) = 1$;
   Our goal is to choose $p(x_i)$ to maximize $H[p]$, so we use a Lagrange multiplier to solve the problem,$\widetilde{H} = -\sum_i p(x_i) ln\, p(x_i) + \lambda(\sum_i p(x_i) - 1)$;
   Finally get $p(x_i) = \frac{1}{M}$, where M is the state number of X.

2. For continuous variable X, $H[x] = -\int p(x) ln\, p(x) dx, \int_{-\infty}^{\infty} p(x) dx = 1, \int_{-\infty}^{\infty} xp(x) dx = \mu, \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2$;
   We use a Lagrange multiplier to solve the maximium problem,$\widetilde{H} = -\int p(x) ln\, p(x) dx + \lambda_1(\int_{-\infty}^{\infty} p(x) dx - 1) + \lambda_2(\int_{-\infty}^{\infty} xp(x) dx - \mu) + \lambda_3(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx - \sigma^2)$;
   Finally get $p(x) = exp\{-1 + \lambda_1 + \lambda_2 + \lambda_3 (x-\mu)^2\}$.?Then $p(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$.

- Attributes of entroy
  1. $H[\boldsymbol{y}|\boldsymbol{x}] = -\iint p(\boldsymbol{x}, \boldsymbol{y}) ln\, p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} d\boldsymbol{x}$. No matter what the target is, the left in the integral is the joint distribution of variables in the target.
  2. $H[\boldsymbol{x}, \boldsymbol{y}] = H[\boldsymbol{y}|\boldsymbol{x}] + H[\boldsymbol{x}]$

- Kullback-Leibler divergence
  1. $p(\boldsymbol{x})$ is the unknown distribution, $q(\boldsymbol{x})$ is the approximating distribution for $p(\boldsymbol{x})$,definite

     $$KL(p\|q) = -\int p(\boldsymbol{x}) ln q(\boldsymbol{x}) d\boldsymbol{x} - (-\int p(\boldsymbol{x}) ln\, p(\boldsymbol{x}) d\boldsymbol{x}) = -\int p(\boldsymbol{x}) ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}$$

     $KL(p\|q)$ is average addtional amount of information required to specify the value $\boldsymbol{x}$(assuming we choose an efficient coding scheme) as a result of using $q(\boldsymbol{x})$ instead of $p(\boldsymbol{x})$

  2. *Jensen's inequality*: Convex function satisfies $f(\sum_{i=1}^{M} \lambda_i x_i) \leq \sum_{i=1}^{M} \lambda_i f(x_i)$,
     where $\lambda_i \geq 0, \sum_i \lambda_i = 1$. $f(\sum_{i=1}^{M} \lambda_i x_i)$ is the point on the function, $\sum_{i=1}^{M} \lambda_i f(x_i)$ is the point on the chord.
  3. Using Jensen's inequality to show $KL(p\|q) \geq 0$:

     $$KL(p\|q) = -\int p(\boldsymbol{x}) ln\{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\} d\boldsymbol{x} \geq -ln\{\int p(\boldsymbol{x}) \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}\} = -ln \int q(\boldsymbol{x}) d\boldsymbol{x} = 0$$

4. Consider how to get $KL(p\|q)$ in the real world, the $p(\boldsymbol{x})$ is estimated by the ? finite observed data,

$$KL(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\{-lnq(\boldsymbol{x}_n|\boldsymbol{\theta})+ln\, p(\boldsymbol{x}_n)\} = \sum_{m=1}^{M} p(\boldsymbol{x}_m)\{-lnq(\boldsymbol{x}_m|\boldsymbol{\theta})+ln\, p(\boldsymbol{x}_m)\}$$

- *Mutual information*:

$$I[\boldsymbol{x},\boldsymbol{y}] = KL(p(\boldsymbol{x},\boldsymbol{y})\|p(\boldsymbol{x})p(\boldsymbol{y})) = -\iint p(\boldsymbol{x},\boldsymbol{y})ln(\frac{p(\boldsymbol{x})p(\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{y})})d\boldsymbol{x}d\boldsymbol{y}$$

It's a method to judge wether $\boldsymbol{x},\boldsymbol{y}$ are 'close' to being independent by considering the Kullback-Leibler divergence between the joint distributionand the product of the marginals.Using the sum and product rules of probability, we can see that

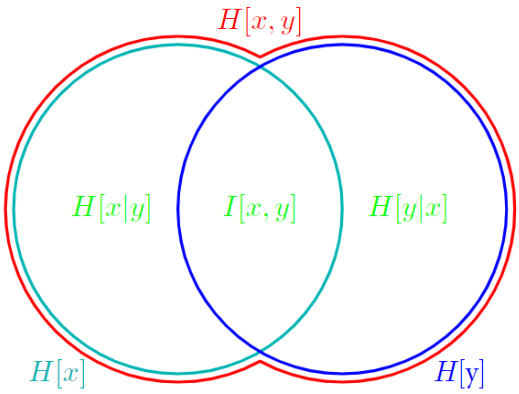

Figure 1: Relaitonship between mutual information and entroy