PRML Note
C01 Introduction

Yang Zhao

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 $\begin{cases} \text{PR problem easy to solve} \\ \text{speed up computation} \end{cases}$
- supervised learning $\begin{cases} \text{classification} \\ \text{regression} \end{cases}$

  unsupervised learning $\begin{cases} \text{clustring} \\ \text{density estimation} \\ \text{dimensions reduction} \end{cases}$

1 Example: Polynomial Curve Fitting

- Linear Model:

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1}$$

  $\boldsymbol{w}$ determined by

$$argmin_{\boldsymbol{w}^*} E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2$$

  which is called error funtion
  M: model comparison or model selection

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

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

- Large value of M $\rightarrow$ Polynomials flexible $\rightarrow$ increasing tuned to the random noises on target values

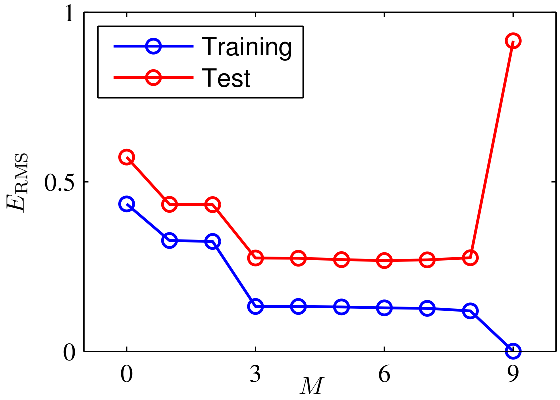

Figure 1: Graphs of the RMS error evalutated on the training set and on an independent test set for various values of M

- 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:

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

  1. shrinkage in statistics
  2. ridge regression
  3. weight decay in NN

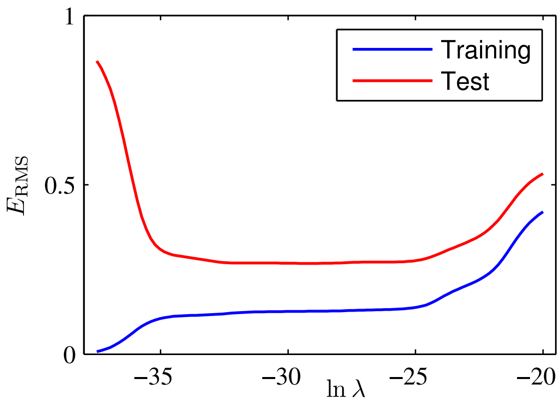

Figure 2: Graphs of the RMS error evalutated on the training set and on an independent test set for $ln\lambda$

- $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^T\boldsymbol{w} = w_0^2 + w_1^2 + \cdots + w_M^2$

1. w_0^2 is often omitted. if not, the result will depend on the choice of the origin for the target variable
2. $\frac{\lambda_1}{2}w_0^2 + \frac{\lambda_2}{2}(w_1^2 + \cdots + w_M^2)$ is also OK

- data set $\begin{cases} \text{training set} \to \boldsymbol{w} \\ \text{validation set(hold-out set)} \\ \to M, \lambda \text{ which is too wasteful of data set} \\ \text{test set} \end{cases}$

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 $p(\boldsymbol{w})$
  Observed data $D = \{t_1, t_2, \cdots, t_N\}$ is expressed through the conditional probability $p(D|\boldsymbol{w})$

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

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

- A data set of observations $X = (x_1, x_2, \cdots, x_N)^T$,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

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

- find the μ and σ^2 to max this likelihood function and the maximum likelihood solution is
$\begin{cases} \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \\ \sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \end{cases}$
and the expectation of it is
$\begin{cases} \mathbb{E}\{\mu_{ML}\} = \mu \\ \mathbb{E}\{\sigma^2_{ML}\} = (\frac{N-1}{N})\sigma^2 \end{cases}$
so

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

- bootstrap: suppose data set consists of N data points $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, we can drawing N data points at random from $\boldsymbol{X}$ to create a new data set $\boldsymbol{X}_B$. This process can be repeated L times

- maximum posterior(MAP): suppose training data comprising N input values $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)^T$ and their corresponding target values $\boldsymbol{t} = (t_1, t_2, \cdots, t_N)^T$. We have

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

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

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

and the log likelihood function is

$$lnp(\boldsymbol{t}|\boldsymbol{x},\boldsymbol{w},\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w}) - t_n\}^2 + \frac{N}{2}ln\beta - \frac{N}{2}ln2\pi$$

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

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

then the posterior distribution for $\boldsymbol{w}$ is

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

and we should maximize the log posterior function

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

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

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

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

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

  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. $p(C_1|x) > p(C_2|x)$, $x \to C_1$

- For k-classes, maximize

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

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

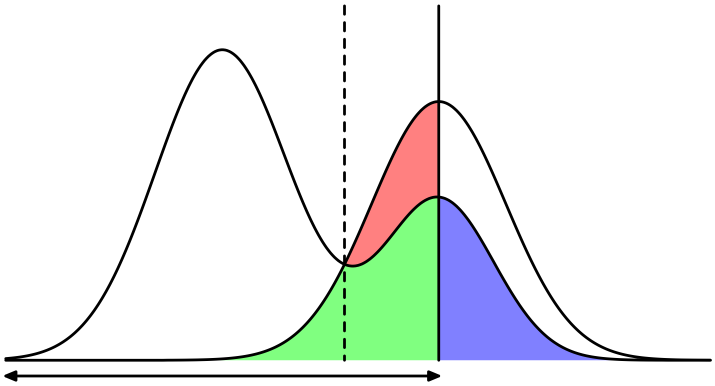

Figure 3: The joint probability of two classes

- 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

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

  and

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

  then normalize to ensure

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

  Explaination:

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

- Combining models: $p(\boldsymbol{x_I}, \boldsymbol{x_B}|C_k) = p(\boldsymbol{x_I}|C_k)p(\boldsymbol{x_B}|C_k)$, so we can get

$$
\begin{aligned}
p(C_k|\boldsymbol{x_I}, \boldsymbol{x_B}) &\propto p(\boldsymbol{x_I}, \boldsymbol{x_B}|C_k)p(C_k) \\
&\propto p(\boldsymbol{x_I}|C_k)p(\boldsymbol{x_B}|C_k)p(C_k) \\
&\propto \frac{p(C_k|\boldsymbol{x_I})p(C_k|\boldsymbol{x_B})}{p(C_k)}
\end{aligned}
$$

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

$$
\frac{\partial J(y)}{\partial y} = \frac{\partial \mathcal{L}}{\partial f} - \frac{d}{dx}\frac{\partial \mathcal{L}}{\partial f'} \tag{3}
$$

  which is called Euler-Lagrange Equation.

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

$$
\mathbb{E}[L] = \iint \mathcal{L}(t, y(\boldsymbol{x}))p(\boldsymbol{x}, t)d\boldsymbol{x}dt \tag{4}
$$

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

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

  So we have

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

- from equation (4), we can rewrite to

$$
\begin{aligned}
\mathbb{E}[L] &= \iint \{y(\boldsymbol{x} - \mathbb{E}[t|\boldsymbol{x}] + \mathbb{E}[t|\boldsymbol{x}] - t)\}^2 p(\boldsymbol{x}, t)d\boldsymbol{x}dt \\
&= \int \{y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}]\}^2 p(\boldsymbol{x})d\boldsymbol{x} + \int Var[t|\boldsymbol{x}]p(\boldsymbol{x})d\boldsymbol{x}
\end{aligned} \tag{5}
$$

  The process to get the equation (5) is below:

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

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

4.

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

5.

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

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 $|y(\boldsymbol{x}) - t|^q$

6 Information Theory

- $h(x) = -log_2 p(x)$
- the entropy of the random variable x(bit): $H[x] = -\sum_x p(x)log_2 p(x)$ and the entropy of the random variable x(nat): $H[x] = -\frac{1}{ln2}\sum_x p(x)lnp(x)$
- if $p(x) = 0$, we have $p(x)log_2 p(x) = 0$
- The conditional entropy of y given x:

$$H[y|x] = -\iint p(y, x)lnp(y|x)dydx$$

  and $H[x, y] = H[y|x] + H[x]$

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

$$\begin{aligned}
KL(p\|q) &= -\int p(x)lnq(x)dx - (-\int p(x)lnp(x)dx) \\
&= -\int p(x)ln\Big\{\frac{q(x)}{p(x)}\Big\}dx
\end{aligned} \tag{6}$$

  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

$$\begin{aligned} I[x,y] &= KL(p(x,y)\|p(x)p(y)) \\ &= -\iint p(x,y)ln\Big\{\frac{p(x)p(y)}{p(x,y)}\Big\}dxdy \end{aligned} \tag{7}$$

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

7 Appendix

- To get the Euler-Lagrange equation3, consider the function

$$J(y) = \int_{x_1}^{x_2} L(x, y(x), y'(x))dx$$

where x_1, x_2 are constants, $y(x)$ is twice continuously differentiable, $y'(x) = dy/dx$, $L(x, y(x), y'(x))$ is twice continuously differentiable with respect to its arguments x, y, y'. If the funtion $J(y)$ attains a local minimum at f, and η is an arbitrary function that has at least one derivative and vanishes at the endpoints x_1 and x_2, then for any number ϵ close to 0,

$$J(f) \leq J(f + \epsilon\eta)$$

Let $\Phi(\epsilon) = J(f + \epsilon\eta)$, and the function $\Phi(\epsilon)$ has a minimum at $\epsilon = 0$, thus

$$\Phi'(0) = \frac{d\Phi}{d\epsilon}\Big|_{\epsilon=0} = \int_{x_1}^{x_2} \frac{dL}{d\epsilon}\Big|_{\epsilon=0} dx = 0$$

Using the total derivative of $L(x, y, y')$

$$\frac{dL}{d\epsilon} = \frac{\partial L}{\partial y}\frac{dy}{d\epsilon} + \frac{\partial L}{\partial y'}\frac{dy'}{d\epsilon} = \frac{\partial L}{\partial y}\eta + \frac{\partial L}{\partial y'}\eta'$$

So,

$$\frac{dL}{d\epsilon}\Big|_{\epsilon=0} = \frac{\partial L}{\partial f}\eta + \frac{\partial L}{\partial f'}\eta'$$

$\eta = 0$ at x_1 and x_2 by definition, therefore,

$$\begin{aligned} \int_{x_1}^{x_2} \frac{dL}{d\epsilon}\Big|_{\epsilon=0} dx &= \int_{x_1}^{x_2} \Big(\frac{\partial L}{\partial f}\eta + \frac{\partial L}{\partial f'}\eta'\Big)dx \\ &= \int_{x_1}^{x_2} \Big(\frac{\partial L}{\partial f}\eta + \frac{d}{dx}\Big(\frac{\partial L}{\partial f'}\eta\Big) - \eta\frac{d}{dx}\frac{\partial L}{\partial f'}\Big)dx \\ &= \int_{x_1}^{x_2} \Big(\frac{\partial L}{\partial f}\eta - \eta\frac{d}{dx}\frac{\partial L}{\partial f'}\Big)dx + \frac{\partial L}{\partial f'}\eta\Big|_{x_1}^{x_2} \\ &= \int_{x_1}^{x_2} \eta\Big(\frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'}\Big)dx = 0 \end{aligned}$$

For η is the arbitrary function, so we have

$$\frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'} = 0 \tag{8}$$