Lecture 4: Supervised Learning: Regression

Qinfeng (Javen) Shi

20 August 2015

Intro. to Stats. Machine Learning
COMP SCI 4401/7401

Table of Contents I

Regression v.s. other types of Supervised Learning

In Supervised Learning, we have (input, correct output) in the training data, *i.e.* Input-output data pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

Based on the output y_i (**not the input**), it breaks down to:

- Classification (discrete output)[1]
- **Regression** (continuous or real valued output)

[1] Novelty detection can also be considered as classification

Example of house price

Classification: Selling a house will make a profit or not ?(yes/no)

Example of house price

Classification: Selling a house will make a profit or not ?(yes/no)

Regression: Sale price for a house? (dollar amount)

Example of house price

Classification: Selling a house will make a profit or not ?(yes/no)

Regression: Sale price for a house? (dollar amount)

Input: $\mathbf{x}$ =

building size	250 sq meters
land size	400 sq meters
# bedrooms	3
# bathrooms	2
# parking	2 (double garage)
# stories	1 (i.e. single story)
...	...

Example of house price

Classification: Selling a house will make a profit or not ?(yes/no)

Regression: Sale price for a house? (dollar amount)

Input: $\mathbf{x} =$

building size	250 sq meters
land size	400 sq meters
# bedrooms	3
# bathrooms	2
# parking	2 (double garage)
# stories	1 (i.e. single story)
...	...

Linear regression output: $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^\top \mathbf{x}$.

Example of house price

Classification: Selling a house will make a profit or not ?(yes/no)

Regression: Sale price for a house? (dollar amount)

Input: $\mathbf{x} =$

building size	250 sq meters
land size	400 sq meters
# bedrooms	3
# bathrooms	2
# parking	2 (double garage)
# stories	1 (i.e. single story)
...	...

Linear regression output: $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^\top \mathbf{x}$.

Reminder: for $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$, $\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^\top \mathbf{x} = \sum_{i=1}^d w^i x^i$.

Example of house price

Classification: Selling a house will make a profit or not ?(yes/no)

Regression: Sale price for a house? (dollar amount)

Input: $\mathbf{x} =$

building size	250 sq meters
land size	400 sq meters
# bedrooms	3
# bathrooms	2
# parking	2 (double garage)
# stories	1 (i.e. single story)
...	...

Linear regression output: $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^\top \mathbf{x}$.

Reminder: for $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$, $\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^\top \mathbf{x} = \sum_{i=1}^d w^i x^i$.

Why called regression, why called linear?

What data do you have?

Historical sales record:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$$

often written as $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$, where

$y_j \in \mathbb{R}$ is the price for the house $\mathbf{x}_j$ (after it is sold).

What data do you have?

Historical sales record:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$$

often written as $\{(\mathbf{x}_j, y_j)\}_{j=1}^{n}$, where

$y_j \in \mathbb{R}$ is the price for the house $\mathbf{x}_j$ (after it is sold).

Regression tries to find $h(\mathbf{x})$ that is as close to y as possible (you only know y after the house is sold).

How to do it?

How to do it?

- Linear Regression
- Ridge Regression
- LASSO
- ...

How to do it?

- **Linear Regression**
- Ridge Regression
- LASSO
- ...

How to measure the error

Goal: to find $h(\mathbf{x})$ that is as close to y as possible.

[2] Squared error turns out to a choice of the loss function in ERM

How to measure the error

Goal: to find $h(\mathbf{x})$ that is as close to y as possible.

Question: How to measure closeness? *i.e.* How well does $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ approximate y?

[2] Squared error turns out to a choice of the loss function in ERM

How to measure the error

Goal: to find $h(\mathbf{x})$ that is as close to y as possible.

Question: How to measure closeness? *i.e.* How well does $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ approximate y?

In **Linear Regression**, we use **squared error** $(h(\mathbf{x}) - y)^2$ for each data point[2].

[2] Squared error turns out to a choice of the loss function in ERM

How to measure the error

Goal: to find $h(\mathbf{x})$ that is as close to y as possible.

Question: How to measure closeness? *i.e.* How well does $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ approximate y?

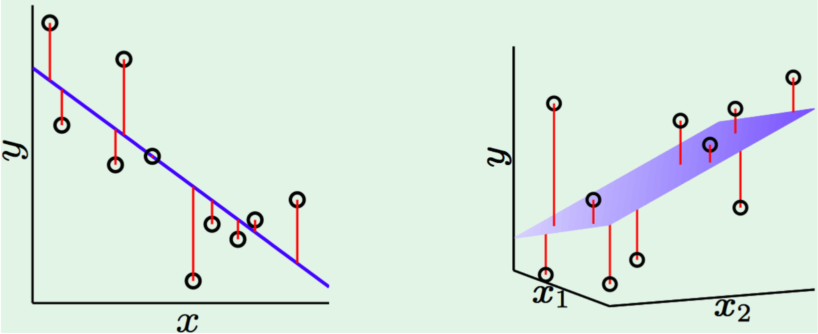
In **Linear Regression**, we use **squared error** $(h(\mathbf{x}) - y)^2$ for each data point[2].

Training error E_{in} (a.k.a **in-sample error, empirical risk**)

$$E_{in}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(h(\mathbf{x}_i) - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

[2]Squared error turns out to a choice of the loss function in ERM

Illustration of linear regression



Matrix expression for empirical risk

$$E_{in}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2 \qquad (1)$$

$$= \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2 \qquad (2)$$

where

$$X = \begin{bmatrix} - \mathbf{x}_1^\top - \\ - \mathbf{x}_2^\top - \\ \vdots \\ - \mathbf{x}_n^\top - \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Minimising empirical risk

$$E_{in}(\mathbf{w}) = \frac{1}{n} \| X\mathbf{w} - \mathbf{y} \|^2$$

Minimising empirical risk

$$E_{in}(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$$

$$\frac{\partial E_{in}(\mathbf{w})}{\partial \mathbf{w}} = \frac{2}{n}X^\top(X\mathbf{w} - \mathbf{y})$$

Minimising empirical risk

$$E_{in}(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$$

$$\frac{\partial E_{in}(\mathbf{w})}{\partial \mathbf{w}} = \frac{2}{n}X^\top(X\mathbf{w} - \mathbf{y}) = 0$$

Minimising empirical risk

$$E_{in}(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$$

$$\frac{\partial E_{in}(\mathbf{w})}{\partial \mathbf{w}} = \frac{2}{n}X^\top(X\mathbf{w} - \mathbf{y}) = 0$$

$$X^\top X\mathbf{w} = X^\top \mathbf{y}$$

Minimising empirical risk

$$E_{in}(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$$

$$\frac{\partial E_{in}(\mathbf{w})}{\partial \mathbf{w}} = \frac{2}{n}X^\top(X\mathbf{w} - \mathbf{y}) = 0$$

$$X^\top X\mathbf{w} = X^\top \mathbf{y}$$

$$\mathbf{w} = X^\dagger \mathbf{y} \quad \text{where} \quad X^\dagger = (X^\top X)^{-1} X^\top$$

Minimising empirical risk

$$E_{in}(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2$$

$$\frac{\partial E_{in}(\mathbf{w})}{\partial \mathbf{w}} = \frac{2}{n}X^\top(X\mathbf{w} - \mathbf{y}) = 0$$

$$X^\top X\mathbf{w} = X^\top \mathbf{y}$$

$$\mathbf{w} = X^\dagger \mathbf{y} \quad \text{where} \quad X^\dagger = (X^\top X)^{-1}X^\top$$

$X^\dagger$ is the 'pseudo-inverse' of X.

Other names and problem

This method is also known as **ordinary least squares (OLS)** or **linear least squares**.

[3] picture from Sakrapee Paisitkriangkrai

Other names and problem

This method is also known as **ordinary least squares (OLS)** or **linear least squares**.

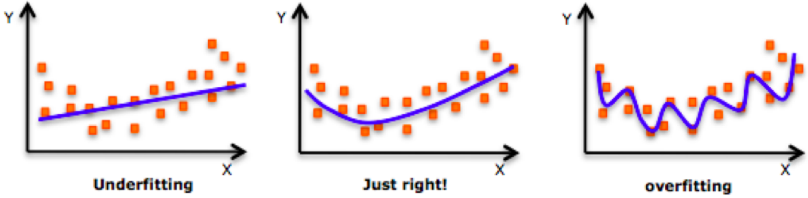
Problem?

[3] picture from Sakrapee Paisitkriangkrai

Other names and problem

This method is also known as **ordinary least squares (OLS)** or **linear least squares**.

Problem?
It may fit the training data too well, and cause **overfitting**[3].



[3] picture from Sakrapee Paisitkriangkrai

Break

Take a break ...

View it in ERM

Recall in Regularised **Empirical Risk Minimisation**,

$$\mathbf{w}_n = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} R_n(\mathbf{w}, \ell) + \lambda \Omega(\mathbf{w}),$$

where ℓ is a loss function, and $\lambda \geq 0$ is the trade-off parameter between the **empirical risk** $R_n(\mathbf{w}, \ell)$ and **regulariser** $\Omega(\mathbf{w})$.

View it in ERM

Recall in Regularised **Empirical Risk Minimisation**,

$$\mathbf{w}_n = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} R_n(\mathbf{w}, \ell) + \lambda \Omega(\mathbf{w}),$$

where ℓ is a loss function, and $\lambda \geq 0$ is the trade-off parameter between the **empirical risk** $R_n(\mathbf{w}, \ell)$ and **regulariser** $\Omega(\mathbf{w})$.

Linear regression (ordinary least squares or linear least squares) is

$$\mathbf{w}_n = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} E_{in}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

To view it in ERM, we simply let $\lambda = 0$, $R_n(\mathbf{w}, \ell) = E_{in}(\mathbf{w})$

View it in ERM

Recall in Regularised **Empirical Risk Minimisation**,

$$\mathbf{w}_n = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}}\, R_n(\mathbf{w}, \ell) + \lambda\Omega(\mathbf{w}),$$

where ℓ is a loss function, and $\lambda \geq 0$ is the trade-off parameter between the **empirical risk** $R_n(\mathbf{w}, \ell)$ and **regulariser** $\Omega(\mathbf{w})$.

Linear regression (ordinary least squares or linear least squares) is

$$\mathbf{w}_n = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}}\, E_{in}(\mathbf{w}) = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

To view it in ERM, we simply let $\lambda = 0$, $R_n(\mathbf{w}, \ell) = E_{in}(\mathbf{w})$, which means $\ell(\mathbf{x}, y, \mathbf{w}) = (\mathbf{w}^\top \mathbf{x} - y)^2$ (squared error/loss) .

Norms and their properties

Regulariser $\Omega(\mathbf{w})$ is often in a form of p-norm.

Definition (p-norm)

Let $p \geq 0$ be a real number. The p-norm of $\mathbf{x} \in \mathbb{R}^d$ is

$$\| \mathbf{x} \|_p := \left(\sum_{j=1}^{d} |x^j|^p \right)^{1/p}$$

Norms and their properties

Regulariser $\Omega(\mathbf{w})$ is often in a form of p-norm.

Definition (p-norm)

Let $p \geq 0$ be a real number. The p-norm of $\mathbf{x} \in \mathbb{R}^d$ is

$$\|\mathbf{x}\|_p := \left(\sum_{j=1}^{d} |x^j|^p\right)^{1/p}$$

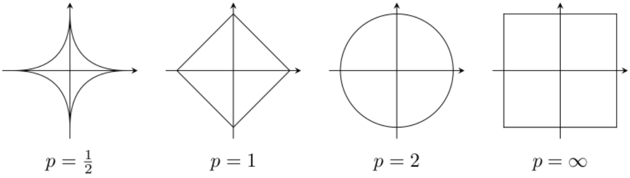


Figure : $\|\mathbf{x}\|_p = 1$

Ridge Regression

What if we keep $R_n(\mathbf{w}, \ell) = E_{in}(\mathbf{w})$, and choose $\lambda > 0$, and regulariser $\Omega(\mathbf{w}) = \| \mathbf{w} \|_2^2$ (like in many other algorithms)?

Ridge Regression

What if we keep $R_n(\mathbf{w}, \ell) = E_{in}(\mathbf{w})$, and choose $\lambda > 0$, and regulariser $\Omega(\mathbf{w}) = \| \mathbf{w} \|_2^2$ (like in many other algorithms)?

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_2^2 \qquad \text{matrix form}$$

Ridge Regression

What if we keep $R_n(\mathbf{w}, \ell) = E_{in}(\mathbf{w})$, and choose $\lambda > 0$, and regulariser $\Omega(\mathbf{w}) = \| \mathbf{w} \|_2^2$ (like in many other algorithms)?

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2 \qquad \text{matrix form}$$

This is called **Ridge Regression**.

Ridge Regression

What if we keep $R_n(\mathbf{w}, \ell) = E_{in}(\mathbf{w})$, and choose $\lambda > 0$, and regulariser $\Omega(\mathbf{w}) = \| \mathbf{w} \|_2^2$ (like in many other algorithms)?

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_2^2 \qquad \text{matrix form}$$

This is called **Ridge Regression**.

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \sum_{j=1}^{d} |w^j|^2 \qquad \text{vector/scalar form}$$

Solving it

Back to the matrix form (much easier to derive the solution)

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2 \qquad \text{matrix form}$$

Taking derivative (w.r.t. $\mathbf{w}$) yields

$$2X^\top(X\mathbf{w} - \mathbf{y}) + 2\mathbf{w}$$

Solving it

Back to the matrix form (much easier to derive the solution)

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2 \qquad \text{matrix form}$$

Taking derivative (w.r.t. $\mathbf{w}$) yields

$$2X^\top(X\mathbf{w} - \mathbf{y}) + 2\mathbf{w} = 0$$

Solving it

Back to the matrix form (much easier to derive the solution)

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2 \qquad \text{matrix form}$$

Taking derivative (w.r.t. **w**) yields

$$2X^\top(X\mathbf{w} - \mathbf{y}) + 2\mathbf{w} = 0$$

The solution becomes:

$$\mathbf{w} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y},$$

where I is the identity matrix (1 on diagonal, 0 else where).

LASSO

What if we replace the ℓ_2 norm with ℓ_1 norm?

LASSO

What if we replace the ℓ_2 norm with ℓ_1 norm?

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$$

This is called **LASSO**.

LASSO

What if we replace the ℓ_2 norm with ℓ_1 norm?

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$$

This is called **LASSO**.

$$\min_{\mathbf{w}} \sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \sum_{j=1}^{d} |w^j| \qquad \text{vector/scalar form}$$

LASSO v.s. Linear Regression, Ridge Regression

$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$	Linear Reg, OLS
$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2$	Ridge
$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$	LASSO

View in constrained form

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$$

Linear Reg, OLS

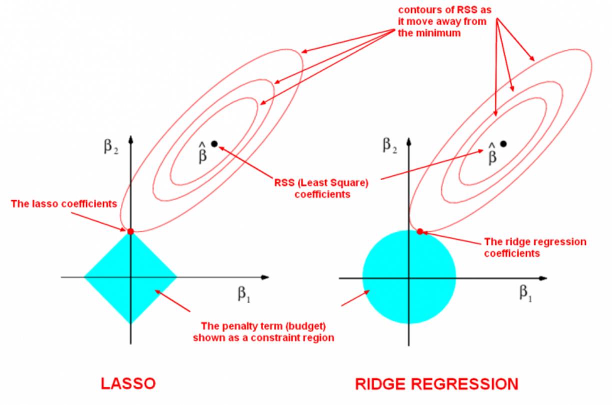
$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$$
$$\text{s.t. } \|\mathbf{w}\|_2 \leq C_1$$

Ridge

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$$
$$\text{s.t. } \|\mathbf{w}\|_1 \leq C_2$$

LASSO

Geometric Interpretation



Sparse solution v.s. dense solution[4].

[4] picture from http://gerardnico.com/wiki/data_mining/lasso