CS722/822: Machine Learning

Instructor: Jiangwen Sun
Computer Science Department

Supervised learning

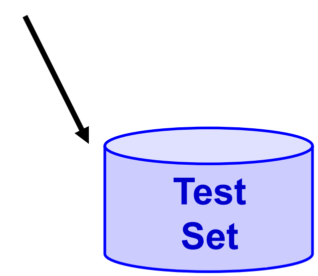
categorical categorical continuous Continuous target

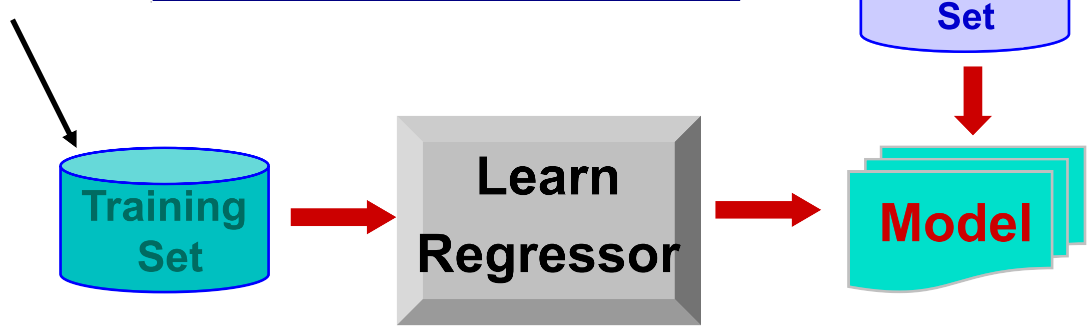
Tid	Refund	Marital Status	Taxable Income	Loss
1	Yes	Single	125K	100
2	No	Married	100K	120
3	No	Single	70K	-200
4	Yes	Married	120K	-300
5	No	Divorced	95K	-400
6	No	Married	60K	-500
7	Yes	Divorced	220K	-190
8	No	Single	85K	300
9	No	Married	75K	-240
10	No	Single	90K	90

Past transaction records, label them

Current data, want to use the model to predict

Refund	Marital Status	Taxable Income	Loss
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



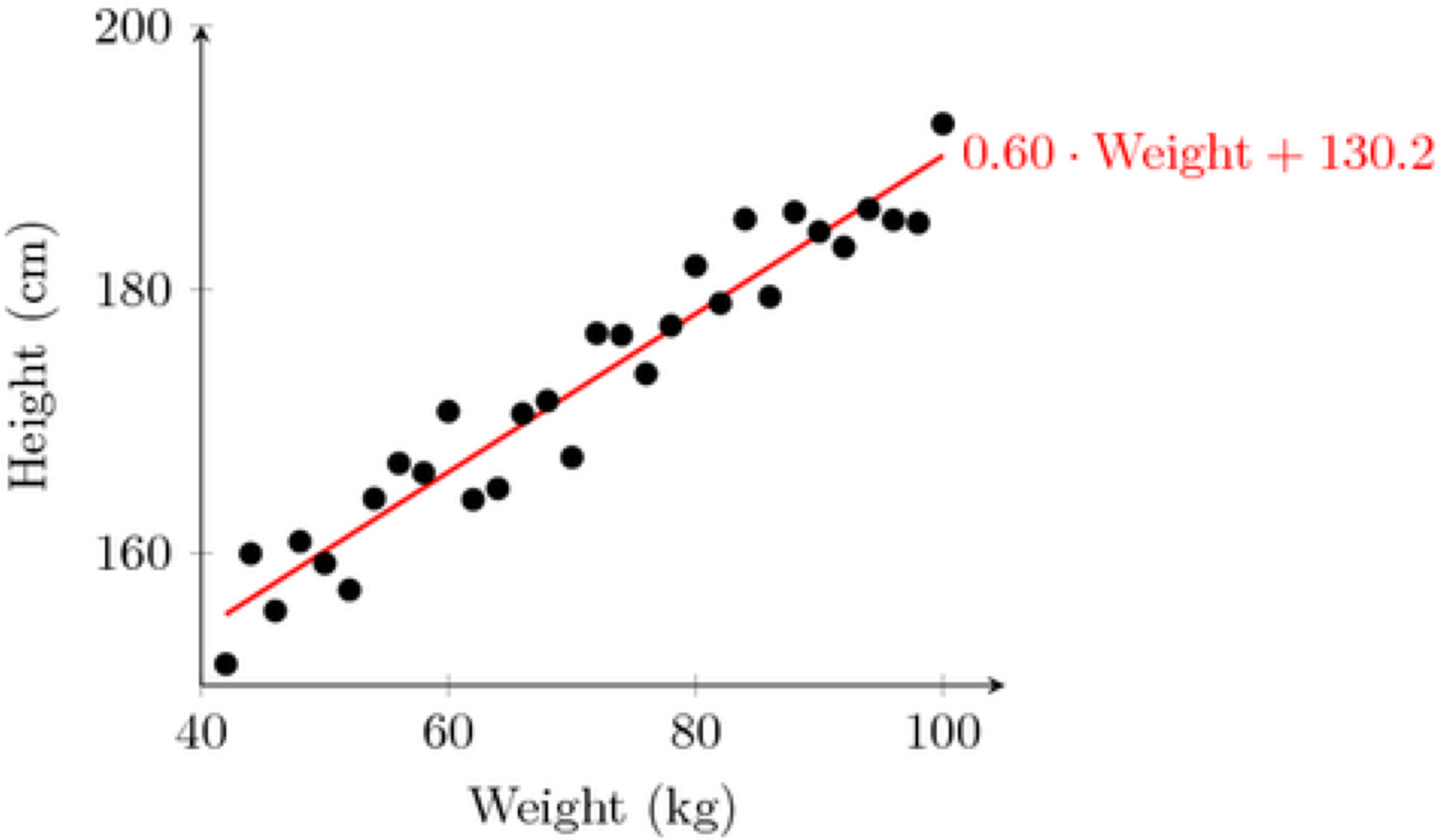


Goals: Predict the possible loss in fraud transaction based on customer records

Regression

- Predict a value of a **real-valued** variable ($\boldsymbol{y}$) based on the values of other variables ($\boldsymbol{X} = \{x_1, \cdots, x_d\}$), assuming a certain model of dependency.
  - In statistics, find a *model* to predict the dependent variable (y) as a function (f) of the values of independent variables (X), mathematically, $y = f(X)$.
- Ultimate Goal: previously unseen examples should be predicted as accurately as possible.
  - A *test set* is used to determine the accuracy of the model, i.e., **testing accuracy**.
  - Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.
  - **Training accuracy** (calculated on training set), not a good measurement for validating the model.

Regression example



Regression method

- Least squares
- Linear regression versus Polynomial regression
- Statistical interpretation of least squares
- Issues: Overfitting versus Underfitting
- Solutions to overfitting:
  - Ridge Regression
  - LASSO – Least Absolute Shrinkage and Selection Operator

Least squares

- Problem: to use some real-valued input variables $X = \{x_1, \cdots, x_d\}$ to predict the value of a target y
- Procedure: We collect training data pairs $(\mathbf{x}_i, y_i)$, $i = 1, \cdots, N$

Tid	**No. Trans**	**Daily Purchase**	**Taxable Income**	**Loss**
1	5	50	125K	**100**
2	8	70	100K	**120**
3	12	17	70K	**-200**
4	15	40	120K	**-300**
5	6	60	95K	**-400**
6	4	44	60K	**-500**
7	16	105	220K	**-190**
8	9	26	85K	**300**
9	2	37	75K	**-240**
10	3	77	90K	**90**

Least squares

- Hypothesis space: suppose we have a model f that maps $\mathbf{x}$ to a value of y
  - f has model parameters $\mathbf{w}$: $f(\mathbf{x}_i; \mathbf{w}) = y_i'$
- Method: minimize the sum of squares:
  - Sum of the squares of the deviation between the observed target value y and the predicted value y'

$$\sum_{i=1}^{N} (y_i - y_i')^2 = \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

Least squares

- Find a function f such that the sum of squares is minimized

$$\min_f \boxed{\sum_{i=1}^{N} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2}$$

Objective/loss function

- Equivalently, find the best $\mathbf{w}$ that minimizes the above squared deviation

$$\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

Least squares

- Linear regression: Least squares with a linear function of parameter $\mathbf{w}$

$$f(\mathbf{x}_i; \mathbf{w}) = \mathbf{x}_i^T \mathbf{w} \quad \longrightarrow \quad \min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^T \mathbf{w}\right)^2$$

- Polynomial regression: Least squares with a polynomial function f (in terms of $\mathbf{x}$, but linear in terms of $\mathbf{w}$)

$$f(\mathbf{x}_i; \mathbf{w}) = \phi(\mathbf{x}_i)^T \mathbf{w} \quad \longrightarrow \quad \min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \phi(\mathbf{x}_i)^T \mathbf{w})^2$$

- Nonlinear regression: Least squares with a non-linear function of parameters $\mathbf{w}$, such as:

$$f(\mathbf{x}_i; \mathbf{w}) = \frac{w_0 x_i}{w_1 + x_i} \quad \longrightarrow \quad \min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \frac{w_0 x_i}{w_1 + x_i}\right)^2$$

Least squares

- Is polynomial regression fundamentally different from linear regression?

Let us discuss

Statistical interpretation of least squares

- Assume that there is a white noise in each observation

$$y_i = f(\mathbf{x}_i; \mathbf{w}) + \varepsilon_i$$

- ε_i follows the standard Gaussian

$$\varepsilon_i \sim \frac{1}{\sqrt{2\pi}} \exp(-\frac{\varepsilon_i^2}{2})$$

- Now, what is the likelihood of observing y_i, $i = 1, \cdots, N$ given $\mathbf{x}_i$

Statistical interpretation of least squares

- The likelihood of observing y_i given $\mathbf{x}_i$ for all $i = 1, \cdots, N$

$$\prod_{i=1}^{N} p(y_i|\mathbf{x}_i;\mathbf{w}) = \prod_{i=1}^{N} p(\varepsilon_i|\mathbf{x}_i;\mathbf{w})$$
$$= \prod_{i=1}^{N} C \exp\left(-\frac{\varepsilon_i^2}{2}\right) = C^N \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \varepsilon_i^2\right)$$
$$= C^N \exp\left(-\frac{1}{2}\sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2\right)$$

- Maximizing the likelihood is equivalent to minimizing the sum of squares

Solve least squares

- Least squares with a linear function of $\mathbf{x}$ and parameters $\mathbf{w}$ is called "linear regression"
- Linear regression has a closed-form solution for w

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^T \mathbf{w}\right)^2$$
$$= \min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$
$$= \min_{\mathbf{w}} E(\mathbf{w})$$

- The minimum is achieved at the zero gradient

The gradient $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$