Linear Model

Wanqi Li

Outline

Two Types of Supervised Learning

- Regression
- Classification

Two LRs

- Linear Regression
- Logistic Regression

Three Interpretations

- Elementary Geometric
- Analytic
- Functional (Probabilistic)

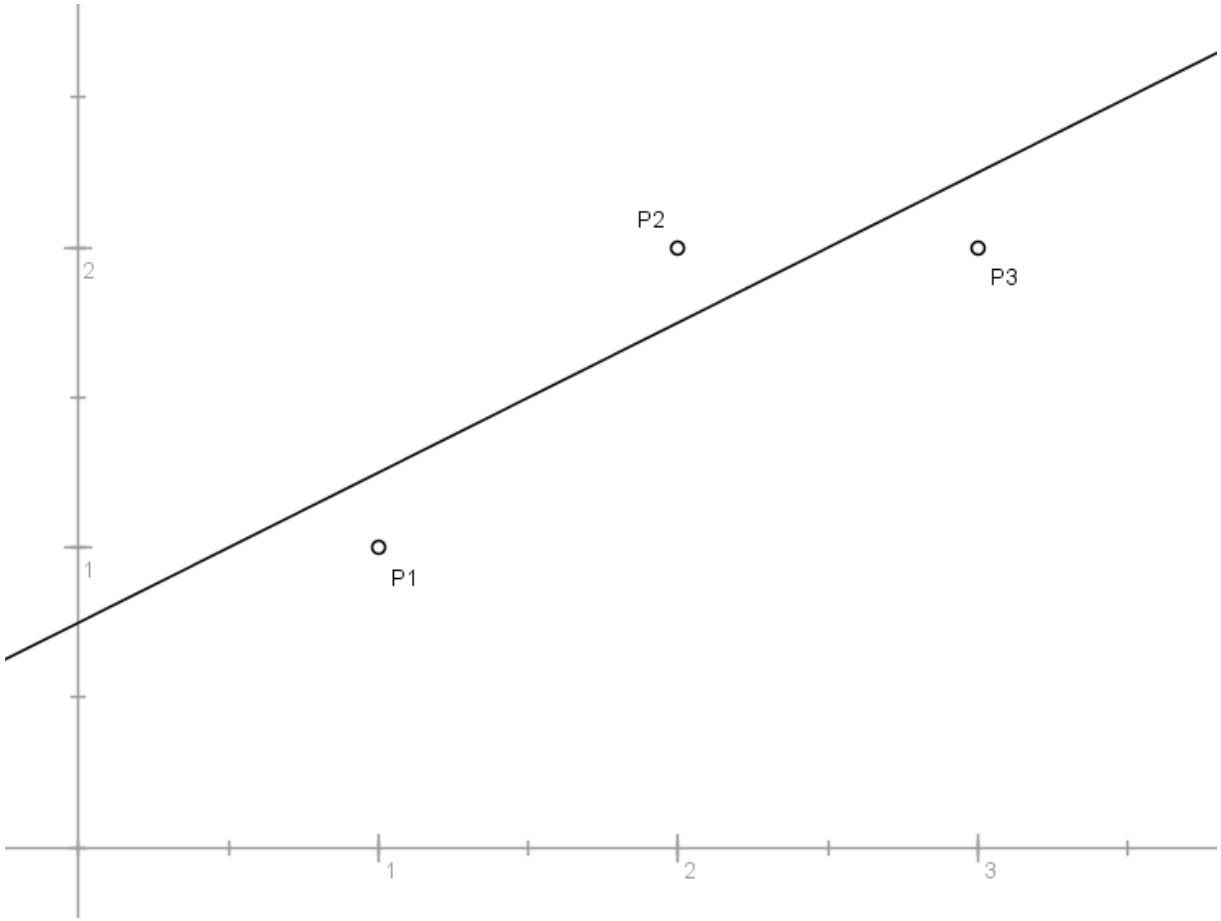
Two Types of Supervised Learning

- Regression
  - Fit a hyperplane* to the training data in $\mathbf{R}^{n+1}$.
  - Minimize the error.
  - Produce a real value.
- Classification
  - Discriminate **tagged** training data with a hyperplane*.
  - Maximize the margin.
  - Produce a discrete value.

* Or curved surface.

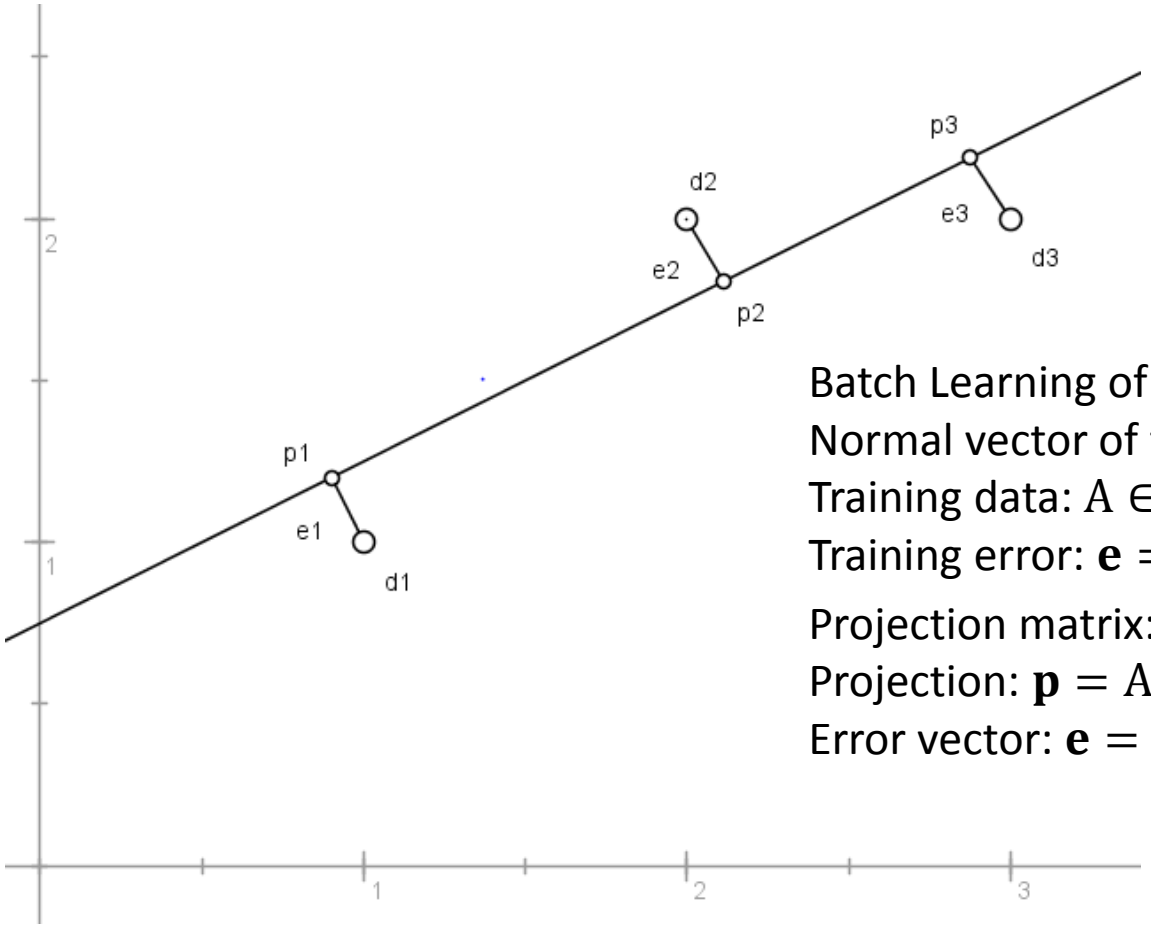
LMS – Elementary Geometric Int.

Fitting a straight line to the sample points.



LMS – Elementary Geometric Int.

Fitting means to minimize the error, which is the distance between each point and its projection on the hyperplane.



Batch Learning of data set A (**Closed Form**):
Normal vector of the regression plane: $\mathbf{x} \in \mathbf{R}^n$
Training data: $\mathrm{A} \in \mathbf{R}^{m*n}, \quad \mathbf{d} \in \mathbf{R}^m$
Training error: $\mathbf{e} = \mathrm{A}\mathbf{x} - \mathbf{d} = (\mathrm{e}_1, \ldots, \mathrm{e}_m)^T$
Projection matrix: $\mathrm{P} = \mathrm{A}(\mathrm{A}^T\mathrm{A})^{-1}\mathrm{A}^T$
Projection: $\mathbf{p} = \mathrm{A}\mathbf{x} = \mathrm{P}\mathbf{d}$
Error vector: $\mathbf{e} = \mathbf{d} - \mathbf{p} = \mathrm{I}\mathbf{d} - \mathrm{P}\mathbf{d} = (\mathrm{I} - \mathrm{P})\mathbf{d}$

LMS – Analytic Interpretation

LMS in Analytic Form (now x is an argument) :

$$\mathbf{x} = \arg\min_{\mathbf{x}\in\mathbf{R}^n} \|\mathbf{e}\|^2$$

$$\|\mathbf{e}\|^2 = \|\mathrm{A}\mathbf{x} - \mathbf{d}\,\|^2 = \sum_{i=1}^{m} \mathrm{e}_i{}^2 = \sum_{i=1}^{m} (\mathbf{x}\boldsymbol{a}_i - d_i)^2 = (\mathrm{A}\mathbf{x} - \mathbf{d})^T (\mathrm{A}\mathbf{x} - \mathbf{d})$$

Different strategies are available to solve this quadratic optimization.

Normal Equations (Looking good but not applicable for large data set):

$$\nabla_{\mathbf{x}} (\mathrm{A}\mathbf{x} - \mathbf{d})^T (\mathrm{A}\mathbf{x} - \mathbf{d}) = 0$$

$$\mathbf{x} = (\mathrm{A}^T \mathrm{A})^{-1} \mathrm{A}^T \mathbf{d}$$

Gradient Descent with learning rate α :

$$\mathrm{x}_i := \mathrm{x}_i - \alpha \frac{\partial}{\partial \mathrm{x}_i} \left(\sum_{i=1}^{m} (\mathbf{x}\boldsymbol{a}_i - d_i)^2 \right) = 0$$

Newton–Raphson method and more numeric methods.

LMS – Probabilistic Int., i.e. ML

Hypothesis (Model): $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$

where all $\epsilon^{(i)}$ are i.i.d. Gaussian distributed. (White Noise)

$$\Pr\left(\epsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

$$\Pr\left(y^{(i)} | \, x^{(i)}, \theta\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

Likelihood: $L(\theta) = \prod_{i=1}^{m} \Pr\left(y^{(i)} | \, x^{(i)}, \theta\right)$

The ML problem:

$$\max_{\theta \in \mathbf{R}^n} \log L(\theta) \quad \Rightarrow \quad \max_{\theta \in \mathbf{R}^n} -(y^{(i)} - \theta^T x^{(i)})^2$$

is equivalent to LMS:

$$\min_{\theta \in \mathbf{R}^n} (y^{(i)} - \theta^T x^{(i)})^2 \Rightarrow \quad \min_{\theta \in \mathbf{R}^n} \|\mathbf{e}\|^2$$

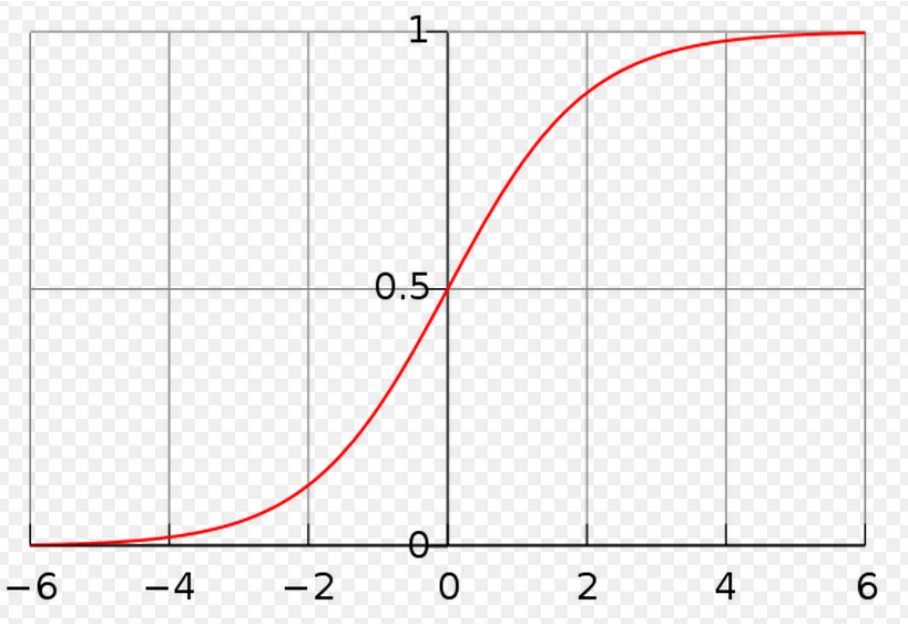
Observations

- Geometry and Analysis describes two different aspects of the same problem.
- Probability Theory is a higher level of abstraction. It enables us to address the same problem using functions (distributions) rather than vectors and quantities. It also provides us an explaination.

Elem. Geometric Int.:	Intuitive, but hard to solve.
Analytic Int.:	Easier to find solution.
Probabilistic Int.:	Start from a model. End up with the same form of Analytic Int.

Logistic Regression

Let's start the derivation with a simple function.



Sigmoid Function:

$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g'(z) = g(z)(1 - g(z))$$

Logistic function is used to model the probability. (Bernoulli Distribution)

$$\Pr(y = 1|x, \theta) = g(\theta^T x)$$

$$\Pr(y = 0|x, \theta) = 1 - g(\theta^T x)$$

Thus the parameter θ can be estimated with ML.

$$L(\theta) = \prod_{i=1}^{m} \Pr\left(y^{(i)} | \, x^{(i)}, \theta\right), \qquad \theta = \arg\max_{\theta \in \mathbf{R}^n} \log L(\theta)$$

Logistic Regression ∉ {Regression}

The process of estimating θ is similar to that of linear regression. But the purpose has changed!

Linear regression: $y = \theta^T x + \epsilon$, a real value.

Logistic Regression: $\Pr(y = 1|x, \theta) = g(\theta^T x)$, the probability.

To predict the discrete value of y, one way is to let

$$y = \begin{cases} 1, \Pr(y = 1|x, \theta) > 0.5 \\ 0, \qquad\qquad \text{Otherwise} \end{cases}$$

So essentially, logistic regression is classification, rather than regression.

More on LR

The probability function of logistic regression belongs to a class of distribution called the **Exponential Family**. So are most of the common distributions.

Distributions in this Exponential Family all belong to the **Generalized Linear Model**.

Maximum Likelihood (ML) applies to all of them.

More on LR

Our previous derivation of ML is called frequentist statistics.

$$L(\theta) = \prod_{i=1}^{m} \Pr\left(y^{(i)} | \, x^{(i)}, \theta\right)$$

In frequentist's point of view, θ is an unknown constant to be estimated.

An alternative of ML is called **Maximum Posteriori** from Bayesian statistics.

$$L(\theta) = \prod_{i=1}^{m} \Pr\left(y^{(i)} | \, x^{(i)}, \theta\right) \Pr(\theta)$$

The prior distribution $\Pr(\theta)$ indicates our "belief" on θ. Some times fabricated priori is used for regularization.

The End