Linear regression

Supervised learning

Data: \( D = \{D_1, D_2, ..., D_n\} \) a set of \( n \) examples
\( D_i = \langle x_i, y_i \rangle \)
\( x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,d}) \) is an input vector of size \( d \)
y_i is the desired output (given by a teacher)

Objective: learn the mapping \( f : X \rightarrow Y \)
s.t. \( y_i \approx f(x_i) \) for all \( i = 1, ..., n \)

- **Regression**: \( Y \) is **continuous**
  Example: earnings, product orders \( \rightarrow \) company stock price
- **Classification**: \( Y \) is **discrete**
  Example: handwritten digit in binary form \( \rightarrow \) digit label
Supervised learning examples

- **Regression**: Y is continuous

Debt/equity | Earnings | Future product orders | Stock price
---|---|---|---
20 | 115 | 20 | 123.45
18 | 120 | 31 | 140.56
...

Data:

Linear regression

- **Function** $f : X \rightarrow Y$
- $Y$ is a linear combination of input components

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \ldots w_d x_d = w_0 + \sum_{j=1}^d w_j x_j$$

$w_0, w_1, \ldots, w_k$ - parameters (weights)
Linear regression

- **Shorter (vector) definition of the model**
  - Include bias constant in the input vector
    \[ x = (1, x_1, x_2, \cdots x_d) \]
    \[ f(x) = w_0x_0 + w_1x_1 + w_2x_2 + \cdots w_dx_d = w^T x \]
  - \( w_0, w_1, \ldots w_k \) - parameters (weights)

### Linear regression. Example

- 1 dimensional input \( x = (x_1) \)
Linear regression. Example.

- 2 dimensional input $\mathbf{x} = (x_1, x_2)$

Linear regression: error

- **Data:** $D_i = \langle x_i, y_i \rangle$
- **Function:** $x_i \rightarrow f(x_i)$
- **Goal:** find the best set of model parameters
- **Error:** a measure of misfit of the model and the data
Linear regression: Error.

• **Data:** \( D_i = \langle x_i, y_i \rangle \)
• **Function:** \( x_i \rightarrow f(x_i) \)
• **Goal:** find the best set of model parameters
• **Error function**
  – a measure of misfit of the model and the data
  – in other words, it measures how much our predictions deviate from the desired answers

**Mean-squared error:** \( Error(w, D) \)

\[
J_n = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, w))^2
\]

• **Learning:**
  We want to find the weights minimizing the error !

---

Linear regression: Optimization.

• We want the weights minimizing the error

\[
J_n = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, w))^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2
\]

• For the optimal set of parameters, derivatives of the error with respect to each parameter must be 0

\[
\frac{\partial}{\partial w_j} J_n(w) = -\frac{2}{n} \sum_{i=1}^{n} (y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,j} = 0
\]

• **Vector of derivatives:**

\[
\text{grad}_w (J_n(w)) = \nabla_w (J_n(w)) = -\frac{2}{n} \sum_{i=1}^{n} (y_i - w^T x_i) x_i = \vec{0}
\]
Linear regression: Optimization.

- $\text{grad}_w (J_n(w)) = \mathbf{0}$ defines a set of equations in $w$

$$\frac{\partial}{\partial w_j} J_n(w) = -2 \sum_{i=1}^{n} (y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,j} = 0$$

$$\frac{\partial}{\partial w_0} J_n(w) = \sum_{i=1}^{n} (y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) = 0$$

$$\frac{\partial}{\partial w_1} J_n(w) = \sum_{i=1}^{n} (y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,1} = 0$$

... 

$$\frac{\partial}{\partial w_d} J_n(w) = \sum_{i=1}^{n} (y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,d} = 0$$

Solving linear regression

$$\frac{\partial}{\partial w_j} J_n(w) = \sum_{i=1}^{n} (y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,j} = 0$$

By rearranging the terms we get a system of linear equations

with $d+1$ unknowns

$$A\mathbf{w} = \mathbf{b}$$

$$w_0 \sum_{i=1}^{n} x_{i,0} 1 + w_1 \sum_{i=1}^{n} x_{i,1} 1 + \ldots + w_j \sum_{i=1}^{n} x_{i,j} 1 + \ldots + w_d \sum_{i=1}^{n} x_{i,d} 1 = \sum_{i=1}^{n} y_i 1$$

$$w_0 \sum_{i=1}^{n} x_{i,0} x_{i,1} + w_1 \sum_{i=1}^{n} x_{i,1} x_{i,1} + \ldots + w_j \sum_{i=1}^{n} x_{i,j} x_{i,j} + \ldots + w_d \sum_{i=1}^{n} x_{i,d} x_{i,j} = \sum_{i=1}^{n} y_i x_{i,j}$$

...
Solving linear regression

• The optimal set of weights satisfies:

\[ \nabla_w (J_n(w)) = -\frac{2}{n} \sum_{i=1}^{n} (y_i - w^T x_i)x_i = 0 \]

Leads to a **system of linear equations (SLE)** with \( d+1 \) unknowns of the form

\[ \mathbf{A} \mathbf{w} = \mathbf{b} \]

Solution to SLE:

\[ \mathbf{w} = \mathbf{A}^{-1} \mathbf{b} \]

Assuming \( \mathbf{X} \) is an \( nxd \) data matrix with rows corresponding to examples and columns to inputs, and \( \mathbf{y} \) is nx1 vector of outputs, then

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

---

Gradient descent solution

**Objective:** optimize the weights in the linear regression model

\[ J_n = Error(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, w))^2 \]

An alternative to SLE solution:

• **Gradient descent**

  **Idea:**
  
  – Adjust weights in the direction that improves the Error
  – The gradient tells us what is the right direction

  \[ \mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_w Error_i(\mathbf{w}) \]

  \[ \alpha > 0 \quad \text{a learning rate} \quad \text{(scales the gradient changes)} \]
**Gradient descent method**

- Descend using the gradient information

\[ \nabla_w \text{Error}(w) \mid_{w^*} \]

- Change the value of \( w \) according to the gradient

\[ w \leftarrow w - \alpha \nabla_w \text{Error}(w) \]

\[ \alpha > 0 \quad \text{a learning rate (scales the gradient changes)} \]

---

**Gradient descent method**

- Iteratively approaches the optimum of the Error function

\[ \text{Error}(w) \]

---
### Batch vs online gradient algorithm

- The error function defined on the complete dataset $D$
  \[
  J_n = Error(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, w))^2
  \]
- We say we are learning the model in the **batch mode**:
  - All examples are available at the time of learning
  - Weights are optimizes with respect to all training examples
- An alternative is to learn the model in the **online mode**
  - Examples are arriving sequentially
  - Model weights are updated after every example
  - If needed examples seen can be forgotten

### Online gradient algorithm

- **The error function** for the complete dataset $D$
  \[
  J_n = Error(w, D) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, w))^2
  \]
- **Error for one example** $D_i = <x_i, y_i>$
  \[
  J_{online} = Error_i(w, x_i) = \frac{1}{2} (y_i - f(x_i, w))^2
  \]
- **Online gradient method: changes weights after every example**
  \[
  w_j \leftarrow w_j - \alpha \frac{\partial}{\partial w_j} Error_i(w)
  \]
- **vector form:**
  \[
  w \leftarrow w - \alpha \nabla_w Error_i(w)
  \]
  \[\alpha > 0\] - Learning rate that depends on the number of updates
Online gradient method

Linear model \[ f(x) = w^T x \]

On-line error \[ J_{online} = Error_j(w) = \frac{1}{2} (y_j - f(x_j, w))^2 \]

On-line algorithm: generates a sequence of online updates

(i)-th update step with: \[ D_i = \langle x_i, y_i \rangle \]

j-th weight:

\[
\begin{align*}
   w_j^{(i)} &\leftarrow w_j^{(i-1)} - \alpha(i) \frac{\partial Error_j(w)}{\partial w_j} |_{w^{(i-1)}} \\
   w_j^{(i)} &\leftarrow w_j^{(i-1)} + \alpha(i)(y_i - f(x_i, w^{(i-1)}))x_{i,j}
\end{align*}
\]

Fixed learning rate: \( \alpha(i) = C \)  
Annealed learning rate: \( \alpha(i) \approx \frac{1}{i} \)

- Use a small constant  
- Gradually rescales changes

Online regression algorithm

Online-linear-regression (stopping Criterion)

Initialize weights \[ w = (w_0, w_1, w_2 \ldots w_d) \]
initialize i=1;

while stoppingCriterion = FALSE
   select the next data point \[ D_i = (x_i, y_i) \]
   set learning rate \( \alpha(i) \)
   update weight vector \[ w \leftarrow w + \alpha(i)(y_i - f(x_i, w))x_i \]
end

return weights

Advantages: very easy to implement, works on continuous data streams
Extensions of simple linear model

Replace inputs to linear units with \( m \) feature (basis) functions to model nonlinearities

\[
f(x) = w_0 + \sum_{j=1}^{m} w_j \phi_j(x)
\]

\( \phi_j(x) \) - an arbitrary function of \( x \)

Original input \( \xrightarrow{\text{New input}} \) Linear model
Extensions of simple linear model

• Models linear in the parameters we want to fit
\[
f(x) = w_0 + \sum_{k=1}^{m} w_k \phi_k(x)
\]

\(w_0, w_1…w_m\) - parameters
\(\phi_1(x), \phi_2(x)…\phi_m(x)\) - feature or basis functions

• Basis functions examples:
  – a higher order polynomial, one-dimensional input \(x = (x_1)\)
  \(\phi_1(x) = x \quad \phi_2(x) = x^2 \quad \phi_3(x) = x^3\)

  – Multidimensional quadratic \(x = (x_1, x_2)\)
  \(\phi_1(x) = x_1 \quad \phi_2(x) = x_1^2 \quad \phi_3(x) = x_2 \quad \phi_4(x) = x_2^2 \quad \phi_5(x) = x_1x_2\)
Extensions of simple linear model

• Models linear in the parameters we want to fit

\[ f(x) = w_0 + \sum_{k=1}^{m} w_k \phi_k(x) \]

\( w_0, w_1, \ldots, w_m \) - parameters
\( \phi_1(x), \phi_2(x), \ldots, \phi_m(x) \) - feature or basis functions

• Basis functions examples:
  - a higher order polynomial, one-dimensional input \( x = (x_1) \)
    \[ \phi_1(x) = x, \quad \phi_2(x) = x^2, \quad \phi_3(x) = x^3 \]
  - Multidimensional quadratic \( x = (x_1, x_2) \)
    \[ \phi_1(x) = x_1, \quad \phi_2(x) = x_1^2, \quad \phi_3(x) = x_2, \quad \phi_4(x) = x_2^2, \quad \phi_5(x) = x_1x_2 \]
  - Other types of basis functions
    \[ \phi_1(x) = \sin x, \quad \phi_2(x) = \cos x \]

• Can be solved exactly like the linear case

---

Extensions of simple linear model

The same techniques as for the linear model to learn the weights

• Error function

\[ J_n = 1/n \sum_{i=1}^{n} (y - f(x_i, w))^2 \]

Assume:
\( \phi(x_i) = (1, \phi_1(x_i), \phi_2(x_i), \ldots, \phi_m(x_i)) \)

\[ \nabla_w J_n(w) = -\frac{2}{n} \sum_{i=1}^{n} (y_i - f(x_i)) \phi(x_i) = 0 \]

• Leads to a system of \( m \) linear equations

\[ w_0 \sum_{i=1}^{n} \phi_j(x_i) + \ldots + w_j \sum_{i=1}^{n} \phi_j(x_i) \phi_j(x_i) + \ldots + w_m \sum_{i=1}^{n} \phi_m(x_i) \phi_j(x_i) = \sum_{i=1}^{n} y_i \phi_j(x_i) \]

• Can be solved exactly like the linear case
Example. Regression with polynomials.

Regression with polynomials of degree m

- **Data instances:** pairs of \( <x, y> \)
- **Feature functions:** \( m \) feature functions
  \[ \phi_i(x) = x^i \quad i = 1, 2, \ldots, m \]
- **Function to learn:**
  \[ f(x, w) = w_0 + \sum_{i=1}^{m} w_i \phi_i(x) = w_0 + \sum_{i=1}^{m} w_i x^i \]

Linear model example
Non-linear (quadratic) model

Linear regression model

- Linear model: \( y = f(x, w) = w^T x \)

- Notice: the above model does not try to explain variation in observed ys for the data
Statistical model of regression

A statistical model of linear regression:

\[ y = \mathbf{w}^T \mathbf{x} + \mathbf{\epsilon} \quad \mathbf{\epsilon} \sim N(0, \sigma^2) \]

\( \mathbf{\epsilon} \) is a random noise, represents things we cannot capture with \( \mathbf{w}^T \mathbf{x} \)

\[ y \sim N(\mathbf{w}^T \mathbf{x}, \sigma^2) \]

• The conditional distribution of \( y \) given \( \mathbf{x} \)

\[
p(y \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma^2} (y - \mathbf{w}^T \mathbf{x})^2 \right]
\]

\[ E(y \mid \mathbf{x}) = \mathbf{w}^T \mathbf{x} \]
ML estimation of the parameters

- **likelihood of predictions** = the probability of observing outputs y in D given w, σ

\[ L(D, w, \sigma) = \prod_{i=1}^{n} p(y_i \mid x_i, w, \sigma) \]

- Maximum likelihood estimation of parameters w
  - parameters maximizing the likelihood of predictions

\[ w^* = \arg \max_w \prod_{i=1}^{n} p(y_i \mid x_i, w, \sigma) \]

- **Log-likelihood** trick for the ML optimization
  - Maximizing the log-likelihood is equivalent to maximizing the likelihood

\[ l(D, w, \sigma) = \log(L(D, w, \sigma)) = \log \prod_{i=1}^{n} p(y_i \mid x_i, w, \sigma) \]

---

ML estimation of the parameters

- **Using conditional density**

\[ p(y \mid x, w, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2} (y - f(x, w))^2] \]

- We can rewrite the log-likelihood as

\[
\begin{align*}
    l(D, w, \sigma) &= \log(L(D, w, \sigma)) = \log \prod_{i=1}^{n} p(y_i \mid x_i, w, \sigma) \\
    &= \sum_{i=1}^{n} \log p(y_i \mid x_i, w, \sigma) \\
    &= \sum_{i=1}^{n} \left[ -\frac{1}{2\sigma^2} (y_i - w^T x_i)^2 - c(\sigma) \right] \\
    &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + C(\sigma)
\end{align*}
\]

Did we see a similar expression before?
ML estimation of the parameters

• Using conditional density
  \[ p(y \mid x, w, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[\frac{1}{2\sigma^2} (y - f(x, w))^2\right] \]

• We can rewrite the log-likelihood as
  \[
  l(D, w, \sigma) = \log(L(D, w, \sigma)) = \log \prod_{i=1}^{n} p(y_i \mid x_i, w, \sigma)
  = \sum_{i=1}^{n} \log p(y_i \mid x_i, w, \sigma)
  = \sum_{i=1}^{n} \left\{ -\frac{1}{2\sigma^2} (y_i - w^T x_i)^2 - c(\sigma) \right\}
  = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + C(\sigma)
  \]

• **Maximizing the predictive log likelihood** with regard to \( w \), is equivalent to minimizing the mean squared error function

---

ML estimation of parameters

• Criteria based on the mean squares error function and the log likelihood of the output are related
  \[
  J_{\text{online}}(y_i, x_i) = \frac{1}{2\sigma^2} \log p(y_i \mid x_i, w, \sigma) + c(\sigma)
  \]

• We know how to optimize parameters \( w \)
  – the same approach as used for the least squares fit

• **But what is the ML estimate of the variance of the noise?**
  • Maximize \( l(D, w, \sigma) \) with respect to variance
    \[
    \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, w^*))^2
    = \text{mean square prediction error for the best predictor}
    \]
**Regularized linear regression**

- If the number of parameters is large relative to the number of data points used to train the model, we face the threat of overfitting (generalization error of the model goes up).
- The prediction accuracy can be often improved by setting some coefficients to zero.
  - Increases the bias, reduces the variance of estimates.
- **Solutions:**
  - Subset selection
  - Ridge regression
  - Lasso regression
  - Principal component regression

**Regularization: motivation**

- If the model is too complex and can cause overfitting, its prediction accuracy can be improved by removing some inputs from the model = setting their coefficients to zero.

\[
f(x) = w_0x_0 + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_dx_d = w^T x
\]

\(w_0, w_1, \ldots, w_k\) - parameters (weights)
Ridge regression

Question: how to force the weights to 0?

• Error function for the standard least squares estimates:

\[
J_n(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2
\]

• We seek:

\[
w^* = \arg \min_w \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2
\]

• Ridge regression:

\[
J_n(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|w\|_{L_2}^2
\]

- Fit to data
- Model complexity penalty

• Where \( \|w\|_{L_2}^2 = \sum_{i=0}^{d} w_i^2 \) and \( \lambda \geq 0 \)

• What does the new error function do?

Ridge regression

Ridge regression:

\[
J_n(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|w\|_{L_2}^2
\]

Term \( \|w\|_{L_2}^2 = \sum_{i=0}^{d} w_i^2 \)

• penalizes non-zero weights with the cost that is proportional to \( \lambda \) (a shrinkage coefficient)

• If an input attribute \( x_j \) has a small effect on improving the error function it is “shut down” by the penalty term

• Inclusion of a shrinkage penalty is often referred to as regularization.

(ridge regression is related to Tikhonov regularization)
**Regularized linear regression**

How to solve the least squares problem if the error function is enriched by the regularization term $\lambda \|w\|^2$?

**Answer:** The solution to the optimal set of weights $w$ is obtained again by solving a set of linear equation.

**Standard linear regression:**

$$\nabla_w (J_n(w)) = -\frac{2}{n} \sum_{i=1}^{n} (y_i - w^T x_i) x_i = 0$$

**Solution:** $w^* = (X^T X)^{-1} X^T y$

where $X$ is an $n \times d$ matrix with rows corresponding to examples and columns to inputs.

**Regularized linear regression:**

$$w^* = (\lambda I + X^T X)^{-1} X^T y$$

---

**Lasso regression**

- **Standard regression:**

  $$J_n(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

- **Lasso regression/regularization:**

  $$J_n(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|w\|_{L1}$$

  Fit to data Model complexity penalty

  $$\|w\|_{L1} = \sum_{i=0}^{d} |w_i|$$

  and $\lambda \geq 0$

- L1 is more aggressive pushing the weights to 0 compared to L2
Lasso vs Ridge penalty

- Lasso (L1) penalty \( \| w \|_{L1} = \sum_{i=0}^{d} |w_i| \)
- Ridge (L2) penalty \( \| w \|_{L2}^2 = \sum_{i=0}^{d} w_i^2 \)

- L1 is more aggressive pushing the weights to 0 compared to L2