Linear Regression

Gradient Descent

- □□**Model**□$y=w\cdot x + b$

  - Model□□□□□□□□□□□□□□□□□□□

    - □□□**Model**□**training data**□□□□□□□□□□**testing data**□□□□□**Overfitting**

- **Loss**□$L(w,b) = \sum_{i=1}^{n}(y-(w \cdot x + b))^2$

- □□□□□□□□□□

  - $w_1 = w_0 - \alpha\frac{\partial L}{\partial w}|_{w=w_0}$
  - $b_1 = b_0 - \alpha\frac{\partial L}{\partial b}|_{b=b_0}$
  - α□□**Learning rate**□□□□□□□□□□

Regularization

- **□□□□Loss function**
  - $L(w,b) = \sum_{i=1}^{n}(y-(w \cdot x + b))^2 + \lambda\sum(w_i)^2$
  - □□λ□□□□□w□**Loss function**□□□
    - λ□□□□□□□□□□□□□□□□w□□□□ 0□□□□□□□□□□□□□
    - λ□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Classification

- ### 1. □□□□□□□
  - □□□□□□□□□□□□□□□□□□

  - □□□□□□□□□□□□□□□□□□$ P(y=k|x)$□□□□□□□□□□□□□□□□□□□

  - □□□□□□□□□□□□□□

    $P(y=k|x)=\frac{P(x | y = k) P(y = k)}{P(x)}$

    □□□

    - $P(y=k|x)$□□□□□□□□□□ x□□□□□ k □□□□
    - $P(x|y=k)$□□□□□□□□ k □□□□ x □□□□
    - $P(y=k)$□□□□□□□□□□□ k □□□□□□□□□
    - $P(x)$□□□□□□□□□□□□□□ 1□□□□□□□□□□□

  ### 2. □□□□□□□□□□

  - **□□□□ $P(y=k|x)$ □□□□ $P(x|y=k)$ □ $P(y=k)$ □□□□□□□**
  - □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□□□□□

  ### 3. □□□□□□

  1. **□□□□□□ $P(y=k)$**□□□□□□□□□□□□□□□□□□□□
  2. **□□□□ $P(x|y=k)$**□
     - □□□□□□□x□□□□□□□□□□□□□□□□□□□□□□□

3. **□□□□□□ $P(y=k|x)$**□□□□□□□□□□□□□□□□□□

---

4. □□□□□□□

- □□□□□□□**MLE**□□□□□□□□□□□μ □□□□□□ Σ□□□□□□□□□□□□□□□□
  - □□$L(\mu,\Sigma) = f_{\mu,\Sigma}(x^1)f_{\mu,\Sigma}(x^2)f_{\mu,\Sigma}(x^3)f_{\mu,\Sigma}(x^4)\dots f_{\mu,\Sigma}(x^{79})$
  - □□□□$L(\mu,\Sigma)$□□□μ,Σ
  - □□□□□□□ $μ1,μ2,Σ\mu^1, \mu^2, \Sigmaμ1,μ2,Σ $□□□□□□□□□□□□□□
- □□□□□□□□□□□□ x□□□ $P(y=1|x)$ □ $P(y=2|x)$□□□□□□□□□□

---

5. □□□□

- $\large P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1)+P(x|C_2)P(C_2)} = \frac{1}{1+\frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}} = \frac{1}{1+e^{-z}} = \sigma(z)$
  - **$\sigma(z)$□sigmoid**□□□□□□□□[0,1]□□□□□□□□□
- □□□z□□$z = wx + b$
  - □□$z = \large ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$□□□□□□□□□
    - $w^T = (\mu^1 - \mu^2)\Sigma^{-1}$
    - $b = -\frac{1}{2}(\mu^1)^T(\Sigma^1)^{-1}\mu^1+\frac{1}{2}(\mu^2)^T(\Sigma^2)^{-1}\mu^2+ln\frac{N_1}{N_2}$
      - □□$\mu^1□\mu^2□\Sigma^1□\Sigma^2□N_1□N_2$□□□□□□□□□□□□□□□□□□□□□□□□□□□□□$\Sigma^1 = \Sigma^2 = \Sigma$

Logistic Regression

Loss function

□□□□□□**$f_{w,b}(X) = P_{w,b}(C_1|x)$**□□□□□□

- □□□□□
□□□□	**x^1**	**x^2**	**x^3**	**$\cdots$**
□□	C_1	C_1	C_2	$\cdots$
□□	$\hat{y} = 1$	$\hat{y} = 1$	$\hat{y} = 0$	$\cdots$
- □□□□□□□$\large f(w,b) = f_{w,b}(x^1)f_{w,b}(x^2)(1-f_{w,b}(x^3))\dots f_{w,b}(x^N)$
- □□□ $f_{w,b}(x)$ □□□□□□□□□ x□□□□□□□□□□□ C_1 □□□□□□

  $\large f_{w,b}(x) = P_{w,b}(C_1 | x)$

  □□□□□□□□□□□□ C_1 □ C_2□□□□□□ C_2 □□□□□□

  $\large P_{w,b}(C_2 | x) = 1 - P_{w,b}(C_1 | x) = 1 - f_{w,b}(x)$

  - □□□**□□w,b□□□□□□□□□$$w^,b^ = argmaxL(w,b) = argmin (- lnL(w,b))$$
  - $$-lnL(w,b) = lnf_{w,b}(x^1)+lnf_{w,b}(x^2)+ln(1-f_{w,b}(x^3)) = \sum_{n}-[\hat{y}^nlnf_{w,b}(x^n)+(1-\hat{y}^n)ln(1-f_{w,b}(x^n))]$$

    - □□**Cross entropy**□$ C(f(x^n),\hat{y}^n) = \sum_{n}-[\hat{y}^nlnf_{w,b}(x^n)+(1-\hat{y}^n)ln(1-f_{w,b}(x^n))]$□□□□□□□□□

□□Gradient Descent□□□□□□

□**Cross entropy**□$C(f(x^n),\hat{y}^n) = \sum_{n}-[\hat{y}^nlnf_{w,b}(x^n)+(1-\hat{y}^n)ln(1-f_{w,b}(x^n))]$□□**Gradient Descent**□

- $\Large\frac{\partial{ln(1-f_{w,b}(x))}}{\partial{w_i}} = \frac{\partial{ln(1-f_{w,b}(x))}}{\partial{z}}\frac{\partial{z}}{\partial{w_i}} = -\sigma(z)x_i$
    - $\large \frac{\partial{ln(1-\sigma(z))}}{\partial{z}} = -\frac{1}{1-\sigma(z)}\sigma(z)(1-\sigma(z))$
    - $\large\frac{\partial{z}}{\partial{w_i}} = x_i$
- □□□$\Large\frac{\partial{lnf_{w,b}(x)}}{\partial{w_i}} = \frac{\partial{lnf_{w,b}(x)}}{\partial{z}}\frac{\partial{z}}{\partial{w_i}} = (1-\sigma(z))x_i$
- □□□$\Large \frac{\partial{(-ln(w,b))}}{w_i} = \sum_n -(\hat{y}^n - f_{w,b}(x^n))x_i^n$
    - □□□□□□□□□□□□□□□□

□□Square Error□□□

- □**Loss function**□□$L(f) = \frac{1}{2}\sum_n(f_{w,b}(x^n)- \hat{y}^n)^2$
- $\large\frac{\partial{(f_{w,b}(x)-\hat{y}^n)^2}}{\partial w_i} = \normalsize 2(f_{w,b}(x) - \hat{y})f_{w,b}(x)(1-f_{w,b}(x))x_i$
    - □$\hat{y}^n = 0$□□□□□□□$f_{w,b}(x) = 1$□□□□□□0□□□□□□□□□□□□□□□□□□□□□□□□□
    - □$\hat{y}^n = 1$□□□□□□□$f_{w,b}(x) = 0$□□□□□□0□□□□□□□□□□□□□□□□□□□□□□□□□

Multi-class Classification

□□□□□□□□□□□weight□bias

- $C_1:w^1,b_1$ $z_1 = w^1 \cdot x+ b_1$
- $C_2:w^2,b_2$ $z_2 = w^2 \cdot x+ b_2$
- $C_3:w^3,b_3$ $z_3 = w^3 \cdot x+ b_3$

1. □□`softmax`□□□□□□

- $f(z) = \large \frac{e^z}{\sum_{i=1}^n e^i}$
- $f(z)$□□□□□[0,1]□□□□□□□□□□□1

2. □□`Cross entropy`□□□□□□□□□□□□□□

- $\large -\sum_{i=1}^n \hat{y_i}lny_i$

Limitation of Logistic Regression

- □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
- □□□□□$\sigma(z) = 0.5$□□$w \cdot x + b = 0$

□□□□□□□Feature transformation□□□□□□□□□□□□□□□

Discriminative and Generative

Discriminative□□□□□□□

□□□□□□□□□□□□□□□□□□□□

- □□□□□□□□□□□□□**x**□□□□□□
- □□□□□$P(y|x)$□□
- □□□□□□□□□□□□□□□□□□□□□□□□□□

Generative□□□□□

□□□□□□□□□□□□□□□□□□□

- □□□□$P(x|y)$□□□□□□□□□□□□□
- □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□w□b
- □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Deep Learning

Three steps for Deep Learning

1. Define a set of funtion

- □□□□□□□□□□□□□□□□
- □□□
  - □□□□□□□□□□□□□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□ ReLU□Sigmoid□Tanh □□
  - □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□Adam□□□□□□□□□□□□□

□□□

- □□□□□□□□□ Sigmoid□□□□□ ReLU□ReLU □□□□□

  ### Sigmoid □□□

  - **□□□ (0, 1)**□□□□
    - □□□□□□□□□□□□□□□□□ 0□□□□□□□
  - □□□□□□□□□□□□□□□□□

  ### ReLU □□□

  - **□□ [0, ∞)**□□□□□□□□□□□□ 0
  - □□□□□□□□□□ sigmoid □□□□□□□□
  - □□□□□□□□□□

  ### □ ReLU □□□

  - **ReLU** □□□□□□□□□□□ 0□□□□□□□□□□□□□□□□□□

2. Goodness of function

- □□□□□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□k□□□□□□□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

3. Pick the best function

- [illegible]
    - [illegible]
    - [illegible]k[illegible]
    - [illegible]
    - [illegible]

Gradient Descent

- Learning Rate[illegible]
    - [illegible]
    - [illegible]

[illegible]**Taylor Series**[illegible]

Adagrad[illegible]

[illegible]

Stochastic Gradient Descent

[illegible]$L=\sum_n(\hat{y^n}-(b + \sum(w_ix_i^n)))^2$

- **Gradient Descent**[illegible]$\theta^i = \theta^{i-1} - \eta \nabla L(\theta^{i-1})$[illegible]
- **Stochastic Gradient Descent**[illegible]**example**[illegible]
    - [illegible]

Feature Scaling

[illegible]

- [illegible]
- [illegible]

[illegible]

- $\large x_i^r \leftarrow \frac{x_i^r - m_i}{\sigma _i}$
    - [illegible]i[illegible]**m_i**[illegible]**σ_i**
    - [illegible]**0**[illegible]**1**

Backpropagation

- [illegible]**Forward pass**[illegible]**Backword pass**[illegible]
    - [illegible]
        - [illegible]
    - [illegible]

image-20250416233353334

Forward pass□□□□□□

Backword pass□□□□□□

BERT

□□□□□□□□□□

Unsupervised Learning

Self-supervised Learning

Word Embedding□□□□□

□□□□□□□□□□

□□□□□

1. □□□□□□□□□□

- □□□□□□□□□□□**(one-hot encoding)**□□□□□□□□□□□□□□□□(□□□□□□□□□□□)
- □□□□□□□□□□□□□□□□□□□□□(□□□50-300□)

2. □□□□□□□

- □□□□□□□□□□□□□□□□□□□□□□□□□□□
- □□□□□□□□□□□□□□□□□□□□□

Predication-based

Prediction-based□□□□□□□ □□□□□□□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□

□□□□

1. □□□□□□□□□□□□□□□□□one-hot encoding□
   - □□□□"□□□□"□□"□"□□□□□"□"□"□□"□□□□□
2. □□□□□□□□□□□□□□□□□□□
   - "□"□□□□□□ × □□□□□ → "□"□□□□

- "□□"□□□□□□ × □□□□□ → "□□"□□□□
  - □□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□□

4. □□□□□□□□□□□□□□□□
   - □□□□□□□□□□□□□□□□□□□

- [illegible]

5. [illegible]
    - [illegible]
    - [illegible]

[illegible]

1. [illegible]
    - [illegible]
    - [illegible]
2. [illegible]
    - [illegible]
    - [illegible]

Seq2Seq

Sequence to Sequence Model[illegible]

[illegible]

Seq2Seq[illegible]

1. [illegible]Encoder[illegible]
2. [illegible]Decoder[illegible]

AT[illegible]Auto-regressive[illegible]VS NAT[illegible]Non-Autoregressive Transformer[illegible]

[illegible]AR/AT[illegible]

- [illegible]
- [illegible]
- [illegible]
- [illegible]GPT[illegible]**Transformer**[illegible]

[illegible]NAT[illegible]

- [illegible]
- [illegible]
- [illegible]
- [illegible]