Iowa State University

Department of Computer Science

COM S 573: Machine Learning

Lecture 4: Probabilistic Interpretation of Linear Regression

Linear Regression: Summary

- Representation

$$\hat{y}_i = \boldsymbol{w}\boldsymbol{x}_i^T$$

Predict a continuous scalar.

- Evaluation

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

- Optimization

$$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w}^* = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

Linear Regression: Probabilistic interpretation

- Evaluation

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

- Residual for each data sample

$$r_i = y_i - \boldsymbol{x}_i^T\boldsymbol{w}$$

What assumptions are we using?

Linear Regression: Assumptions

- Data are linear

$$\hat{y}_i = \boldsymbol{w}^T \boldsymbol{x}_i$$

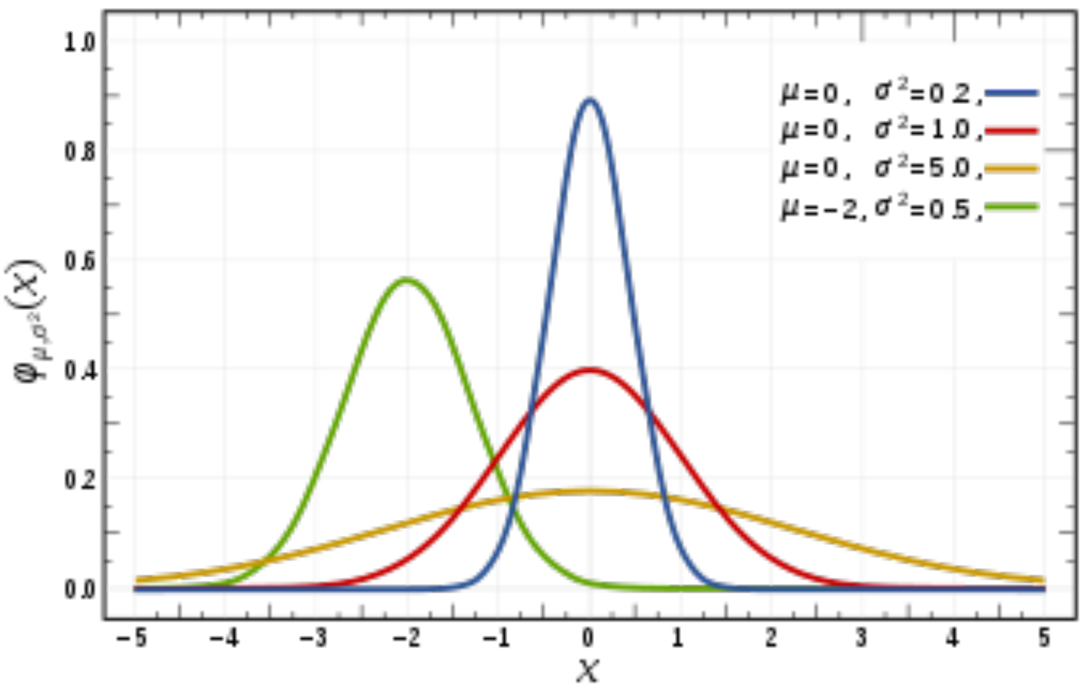
- Residuals are Independent and identically distributed

$$p(r_i, r_j) = p(r_i)p(r_j)$$

- Each residual follows normal distribution

$$r_i \sim \mathcal{N}(\mu, \sigma^2)$$

Linear Regression: Normal Distribution



$$x \sim \mathcal{N}(\mu, \sigma^2) \rightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Linear Regression: likelihood

- Given input samples

$$\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$$

- Likelihood vs. Probability

$$\mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) \equiv p(\boldsymbol{x}|\boldsymbol{w})$$

- Likelihood of all input samples

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{w}) &= p(x_1, x_2, \cdots, x_n|\boldsymbol{w}) \\ &= p(x_1|\boldsymbol{w})p(x_2|\boldsymbol{w})\cdots p(x_n|\boldsymbol{w}) \\ &= \prod_{i=1}^{n} p(x_i|\boldsymbol{w}) \end{aligned}$$

Why?

Linear Regression: Optimization

- Maximum Likelihood estimator (MLE)

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \max_{\boldsymbol{w}} \prod_{i=1}^{n} p(x_i|\boldsymbol{w})$$

- Log-likelihood

$$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \log \prod_{i=1}^{n} p(x_i|\boldsymbol{w}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$$

Linear Regression: Optimization

Why 0?

- Noisy observation model

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + r_i, r_i \sim \mathcal{N}(0, \sigma^2)$$

- Probability

$$p(r) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r-\mu)^2}{2\sigma^2}}$$

- Likelihood

$$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$$

Linear Regression: Likelihood

- $\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$

$$\boxed{p(r) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r-\mu)^2}{2\sigma^2}}}$$

Linear Regression: Likelihood

- $$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$$
  $$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}}$$

$$\boxed{p(r) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r-\mu)^2}{2\sigma^2}}}$$

Linear Regression: Likelihood

- $$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$$
  $$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}}$$
  $$= \sum_{i=1}^{n} -\log\sqrt{2\pi}\sigma - \sum_{i=1}^{n} \frac{(\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}$$

Linear Regression: Likelihood

- $$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$$
$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}}$$
$$= \sum_{i=1}^{n} - \log \sqrt{2\pi}\sigma - \sum_{i=1}^{n} \frac{(\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}$$
$$= -\frac{1}{2} \left(\frac{1}{\sigma^2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2 + n \log \sigma^2 + const \right)$$

Linear Regression: Likelihood

- $$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}}$$

$$= \sum_{i=1}^{n} - \log \sqrt{2\pi}\sigma - \sum_{i=1}^{n} \frac{(\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}$$

$$= -\frac{1}{2} \left(\frac{1}{\sigma^2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{x}_i^T \boldsymbol{w})^2 + n \log \sigma^2 + const \right)$$

$$= -\frac{1}{2} \left(\frac{1}{\sigma^2} RSS(\boldsymbol{w}) + n \log \sigma^2 \right) + const$$

Maximize likelihood = minimize RSS

Linear Regression: Likelihood

- Optimization

$$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = -\frac{1}{2}\left(\frac{1}{\sigma^2} RSS(\boldsymbol{w}) + n \log \sigma^2\right) + const$$

- Find the optimal

$$\boldsymbol{w}^*, \sigma^*$$

HW1

Linear Regression: Summary

- Representation

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + r_i, r_i \sim \mathcal{N}(0, \sigma^2)$$

- Evaluation

$$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = \log \prod_{i=1}^{n} p(x_i|\boldsymbol{w}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{w})$$

- Optimization

$$\log \mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = -\frac{1}{2}\left(\frac{1}{\sigma^2} RSS(\boldsymbol{w}) + n \log \sigma^2\right) + const$$

Linear Models: Next

- Linear regression: predict a scalar
  - House price
  - Weight of a planet
- Linear perceptron: classifier of discrete prediction
  - Predict an animal is a dog or not
  - Predict an image contains a square or not
- Logistic regression: classifier based on a probability
  - Predict how likely a team win
  - Predict how likely tomorrow is sunny