Q & A: Bayesian Learning

Lecture series „Machine Learning“

Niels Landwehr

Research Group „Data Science“
Institute of Computer Science
University of Hildesheim

Quiz: Bayes Rule

- The **Bayes rule** states that for any two random variables u, v it holds that

$$p(u \mid v) = \frac{p(v \mid u)\, p(u)}{p(v)}$$

- Assume you get tested for a rare disease. You know the following probabilities:

$p(test = 1 \mid disease = 1) = 0.9$ # The test has a false-negative rate of 10%

$p(test = 1 \mid disease = 0) = 0.1$ # The test has a false-positive rate of 10%

$p(disease = 1) = 0.01$ # 1% of population have the disease

$p(test = 1) = 0.108$ # 10.8% of population test positive (follows)

- Given that you were tested positive, what is the probability that you have the disease?
  - It is around 20%
  - It is around 8%
  - It is around 1%
  - Cannot say based on the information provided

Solution: Bayes Rule

- The **Bayes rule** states that for any two random variables u, v it holds that

$$p(u \mid v) = \frac{p(v \mid u)\, p(u)}{p(v)}$$

- We can apply the Bayes rule as follows:

$$p(disease = 1 \mid test = 1) = \frac{p(test = 1 \mid disease = 1)\, p(disease = 1)}{p(test = 1)}$$

- Plugging in the numbers, we obtain

$$p(disease = 1 \mid test = 1) = \frac{0.9 \cdot 0.01}{0.108} \approx 0.083$$

- Therefore the probability is around 8%

Quiz: Coin Tosses

- Assume we toss a coin with unknown parameter $\theta \in [0,1]$, and observe as data $\mathcal{D}$ 5 head tosses and 0 tail tosses
- As prior distribution, we assume a Beta distribution as given in the lecture with hyperparameters $\alpha_h = 5,\ \alpha_t = 5$

- According to the posterior distribution, what is the probability that $\theta = 1$?
  - $p(\theta = 1 \mid \mathcal{D}) \approx 0.05$
  - $p(\theta = 1 \mid \mathcal{D}) \approx 0.11$
  - $p(\theta = 1 \mid \mathcal{D}) \approx 0.3$
  - $p(\theta = 1 \mid \mathcal{D}) = 0$
  - $p(\theta = 1 \mid \mathcal{D}) = 1$

Solution: Coin Tosses

- **Solution**: The posterior distribution is again given by a Beta distribution with parameters $\alpha_h + N_h = 10$ and $\alpha_t + N_t = 5$

- The Beta distribution in general is given by

$$p(\theta) = Beta(\theta \mid \alpha_h, \alpha_t)$$
$$= \frac{\Gamma(\alpha_h + \alpha_t)}{\Gamma(\alpha_h)\Gamma(\alpha_t)} \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$$

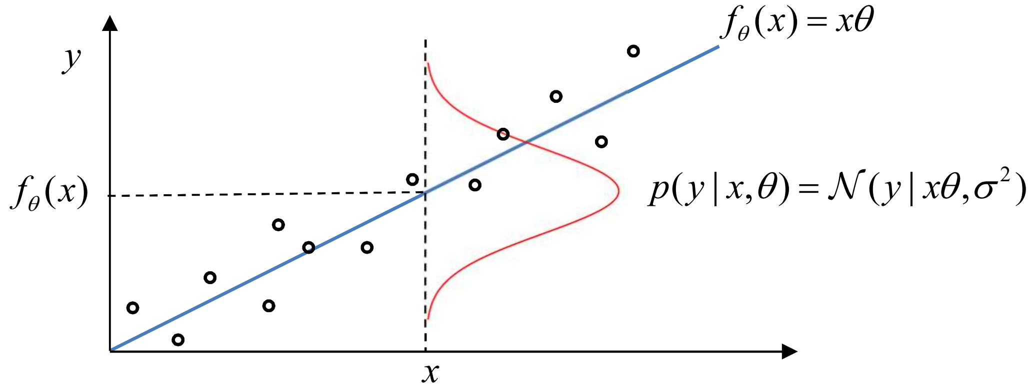
so the posterior in our case is

$$p(\theta \mid \mathcal{D}) = Beta(\theta \mid 10, 5)$$
$$= \frac{\Gamma(15)}{\Gamma(10)\Gamma(5)} \theta^{9}(1-\theta)^{4}$$

- Clear that $p(\theta = 1 \mid \mathcal{D}) = 0$

Quiz: Bayesian Linear Regression

- Assume a univariate Bayesian linear regression model $f_\theta(x) = x\theta$:



- On a given data set $\mathcal{D}$, we train two models: model A with hyperparameter $\sigma = 1$ and model B with hyperparameter $\sigma = 10$. Everything else stays the same.
- What can we say about the variance of the posterior distribution $p(\theta \mid \mathcal{D}) = \mathcal{N}(\theta \mid \mu_{pos}, \sigma^2_{pos})$ of models A and B?
  - Variance of posterior distribution for model B is **higher** than for model A
  - Variance of posterior distribution for model B is **lower** than for model A
  - Variance of posterior distribution is **identical** for both models
  - Cannot say based on the information provided

Solution: Bayesian Linear Regression

- **Solution**: the posterior distribution is given by

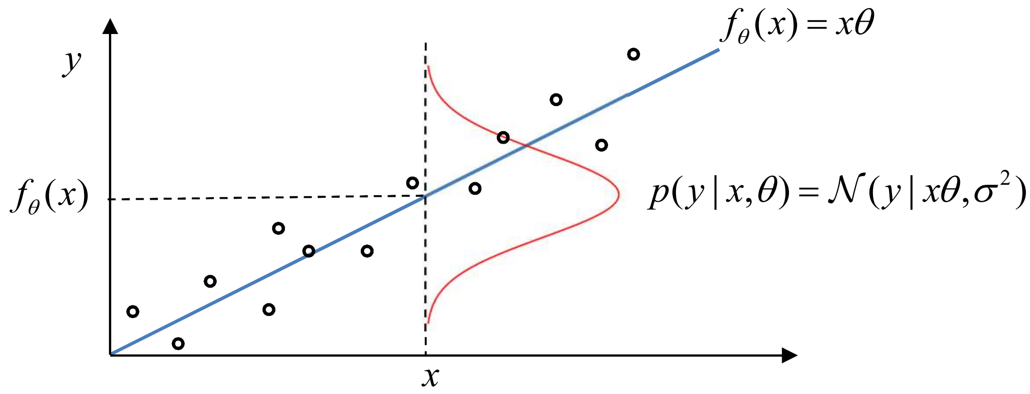
$$p(\theta \mid \mathbf{y}, \mathbf{X}) = \mathcal{N}(\theta \mid \bar{\theta}, A^{-1})$$

$$\text{where } A = \sigma^{-2} \underbrace{\mathbf{X}^{\mathrm{T}} \mathbf{X}}_{\in \mathbb{R}_{>0}} + \sigma_p^{-2} \quad \text{and} \quad \bar{\theta} = \sigma^{-2} A^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y}$$

- As can be seen, increasing the data variance σ leads to smaller A and therefore higher variance

Solution: Bayesian Linear Regression

- **Intuitive Explanation**: increasing the size of the hyperparameter σ corresponds to the assumption that the data fluctuates more strongly around an assumed true model θ^* :



- The more noise/fluctuations there are in the data, the less the data tells us about the true model
- Therefore, at the same data size set, the remaining uncertainty about what the true model is will be higher if σ is higher
- Therefore, the variance of the posterior distribution is higher