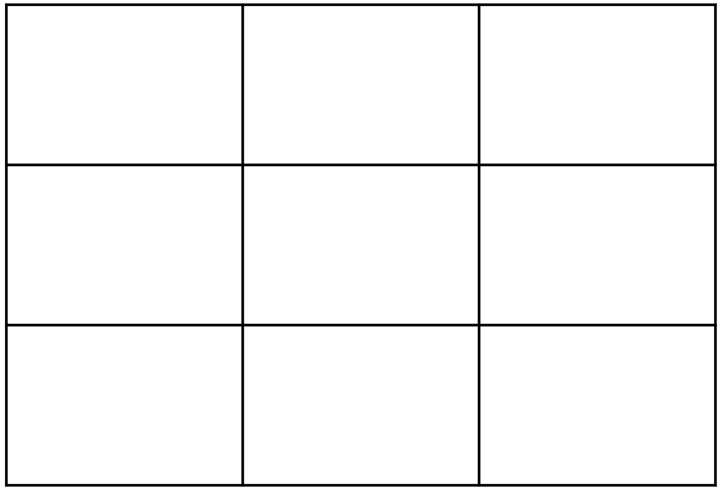
Bayesian Models for Machine Learning

Introduction



Entire space: Ω

A_i is the i-th row

B_i is the i-th column

- We have points lying in space. We pick one of these points uniformly at random.

$$P(x \in A_1) = \frac{\#A_1}{\#\Omega} \qquad\qquad P(x \in B_1) = \frac{\#B_1}{\#\Omega}$$

- Conditional probability: this is the probability that $x \in A_1$ given that I know $x \in B_1$.

$$P(x \in A_1 | x \in B_1) = \frac{\#(A_1 \cap B_1)}{\#B_1} = \frac{\#(A_1 \cap B_1)}{\#\Omega} \frac{\#\Omega}{\#B_1} = \frac{P(x \in A_1 \& \, x \in B_1)}{P(x \in B_1)}$$

Let A and B be two events, then:

$$P(A|B) = \frac{P(A,B)}{P(B)} \qquad\qquad P(A|B)P(B) = P(A,B)$$

$P(A|B)$: conditional probability distribution

$P(A,B)$: joint probability distribution

$P(B)$: marginal probability distribution

$$P(B) = \frac{\#B}{\#\Omega} = \frac{\sum_{i=1}^{3} \#(A_i \cap B)}{\#\Omega} = \sum_{i=1}^{3} \frac{\#(A_i \cap B)}{\#\Omega} = \sum_{i=1}^{3} P(A_i, B)$$

Each A_i has to be disjoint and the union of all A_i has to be equal to the entire space.

Bayes theorem

$$P(A,B) = P(A|B)P(B)$$

By symmetry, we have: $$P(A,B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(A_i,B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

An example will be on board. (medical test)

A person tests ‘positive’ for the disease. What’s the probability he has it?

$$A = \begin{cases} 1 & \text{Person has disease} \\ 0 & \text{Person has no disease} \end{cases} \qquad B = \begin{cases} 1 & \text{Test for disease ‘positive’} \\ 0 & \text{Test for disease ‘negative’} \end{cases}$$

Bayes modeling

Applying Bayes rules to the unknown variables of a data modeling problem is called Bayesian modeling.

Linear regression

Problem setup

- Given a data set of the form $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. The goal is to learn a prediction

rule from this data, so that given a new x^* we can predict its associated unobserved y^*.

Linear regression

$$y_i = x_i^T \omega + \epsilon_i$$

where ω is the parameter and ϵ_i is the noise. We make the assumptions $\epsilon_i \sim N(0, \sigma^2)$

So the likelihood:

$$p(y_i | x_i, \epsilon_i, \omega) \sim N(y_i | x_i^T \omega, \sigma^2)$$

For a Bayesian treatment of linear regression we need a prior probability distribution over model parameters ω:

$$p(\omega) \sim N(\omega | 0, \lambda^{-1})$$

Using Bayes rule, the posterior distribution is:

$$p(\omega | x_i, y_i) \sim N(y_i | x_i^T \omega, \sigma^2) N(\omega | 0, \lambda^{-1} \mathrm{I})$$

Linear regression

$$p(\omega|x_i, y_i) \sim N(y_i|x_i^T\omega, \sigma^2)N(\omega|0, \lambda^{-1}\mathrm{I})$$

$$p(\omega|x_i, y_i) \propto \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i^T\omega)^2\right)\exp(-\frac{\lambda}{2}\omega^T\omega)$$

$$\log p(\omega|x_i, y_i) = -\frac{1}{2\sigma^2}(y_i - x_i^T\omega)^2 - \frac{\lambda}{2}\omega^T\omega$$

Maximizing the log posterior w.r.t. ω gives the maximum-a-posteriori (MAP) estimate of ω.

$$\log p(\omega|x_i, y_i) = -\frac{1}{2\sigma^2}(y_i - x_i^T\omega)^2 - \underline{\frac{\lambda}{2}\omega^T\omega}$$

L2 regularizer

(Ridge regression)

On the other hand, it can be also proven that the posterior distribution is a multivariate Gaussian with mean μ and covariance Σ, where

$$p(\omega|x_i, y_i) \propto \exp(-\frac{1}{2}(\omega - \mu)^T\Sigma^{-1}(\omega - \mu))$$

$\Sigma = (\lambda I + \frac{1}{\sigma^2}\sum_i x_i x_i^T)^{-1}$ $\mu = \Sigma(\frac{1}{\sigma^2}\sum_i y_i x_i)$ (Proof 1)

Linear regression

Posterior predictive distribution

For Bayesian linear regression, we want to predict a new y^* given its associated x^* and all previous observed pairs (x_i, y_i). In other words:

$$p(y^*|x^*, \vec{y}, X) = \int p(y^*|x^*, \omega)\, p(\omega|\vec{y}, X) d\omega$$

- Notice that the predictive distribution can be written as a **marginal distribution** over the model parameters.
- The posterior predictive distribution includes uncertainty about parameters ω into predictions by **weighting** the conditional distribution $p(y^*|x^*, \omega)$ with posterior probability of weights $p(\omega|\vec{y}, X)$ over the entire weight parameter space.
- We can get the **expected value** of y^* at new location x^* as well as the **uncertainty** for that location.

Similarly,
$$p(y^*|x^*,\vec{y},X) = \int p(y^*|x^*,\omega)\,p(\omega|\vec{y},X)d\omega$$
$$= \int (2\pi\sigma^2)^{-1/2}\exp(-\frac{\left(y^*-x^{*T}\omega\right)^2}{2\sigma^2})(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\omega-\mu)^T\Sigma^{-1}(\omega-\mu))$$

After a long derivation,
$$p(y^*|x^*,\vec{y},X) \sim N(x^{*T}\mu, \sigma^2 + x^{*T}\Sigma x^*)$$

Difference between simple linear regression and Bayesian linear regression:

(A specific example will be shown in Jupyter notebook)

Simple linear regression:

- Ordinary Least Squares (OLS);
- Point Estimation (MLE: maximum likelihood estimation);
- One Single set of parameters.

Bayesian linear regression:

- Prior distribution over weights;
- Posterior distribution over weights and outputs;
- Capturing uncertainties