Statistical Techniques in Robotics (16-831, F14) Lecture#16 (Thursday November 6)

Bayesian Linear Regression

Lecturer: Drew Bagnell *Scribe: Rushane Hua, Dheeraj R. Kambam*

1 Bayesian Linear Regression

In the last lecture, we started the topic of Bayesian linear regression. The problem can be represented by the following graphical model:

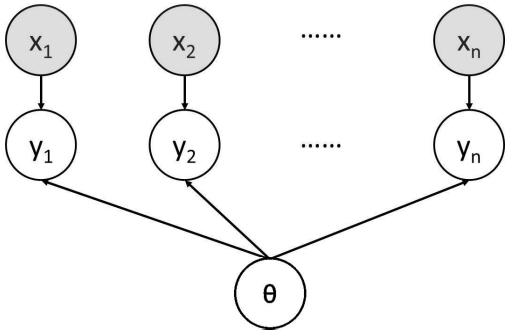


Figure 1: Bayesian linear regression model. x_i's are known.

where

- $x_i \in \mathbb{R}^n$ is the i^{th} set of features in the dataset,
- $y_i \in \mathbb{R}$ is the true outcome given x_i,
- $theta \in \mathbb{R}$ is the parameter vector of the learned model.

The problem we are solving is to find a θ that can make the best prediction on the output $y = \theta^T x$ given an input x.

1.1 Assumption

We assume that the prior distribution of θ is a normal distribution $\mathcal{N}(\mu, \Sigma)$ with mean μ and covariance matrix Σ, and the probability of θ is given by

$$P(\theta) = \frac{1}{Z} \exp\{-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)\}, \tag{1}$$

where $\mu = E_{P(\theta)}[\theta]$ and $\Sigma = E_{P(\theta)}[(\theta - \mu)^T(\theta - \mu)]$.

[0]Some content adapted from previous scribes: Carl Doersch

Equation (1) is called the *moment parametrization* of θ since it consists of the *first moment* (μ) and the *second moment* (Σ, also called the *central moment*) of the variable θ. Z is a normalization factor with the value $\sqrt{(2\pi)^n \det(\Sigma)}$, where n is the dimension of θ. To prove this, one can translate the distribution to center it at the origin, and do change of variables so that the distribution has the form $P(\theta') = \frac{1}{Z}\exp\{-\frac{1}{2}\theta'^T\theta'\}$. Then, express θ' in polar coordinates and integrate over the space to compute Z.

1.2 Prediction

With the Bayesian linear regression model, we would like to know the probability of an output y_{t+1} given an new input x_{t+1} and the set of data $D = \{(x_i, y_i)\}_{i=1,\dots,t}$. To compute the probability $P(y_{t+1}|x_{t+1}, D)$, we introduce θ into this expression and marginalize over it

$$P(y_{t+1}|x_{t+1}, D) = \int_{\theta\in\Theta} P(y_{t+1}|x_{t+1}, \theta, D)P(\theta|x_{t+1}, D) \tag{2}$$

Because D tells no more than what θ does, $P(y_{t+1}|x_{t+1}, \theta, D)$ is essentially $P(y_{t+1}|x_{t+1}, \theta)$. Also, from the graphical model we know that $P(\theta|x_i, D)$ is $P(\theta|D)$ since y_i is known and thus θ and x_i are independent. Now, equation (2) becomes

$$P(y_{t+1}|x_{t+1}, D) = \int_{\theta\in\Theta} P(y_{t+1}|x_{t+1}, \theta)P(\theta|D) \tag{3}$$

Computing (3) is hard with the moment parametrization of normal distributions but not with the natural parametrization.

1.3 Natural Parametrization of Normal Distributions

The normal distribution $P(x) = \frac{1}{Z}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$ can also be expressed as

$$P(x) = \frac{1}{Z}\exp\{J^T x - \frac{1}{2}x^T P x\} \tag{4}$$

The natural parametrization simplifies the multiplication of normal distributions as it becomes addition of the J and P matrices of different distributions.

Transforming the moment parametrization to the natural parametrization can be done by first expanding the exponent:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1} + \mu^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu \tag{5}$$

The last term in equation (5) has nothing to do with x and can therefore be absorbed into the normalizer. By comparing (4) and (5),

$$\begin{aligned} J &= \Sigma^{-1}\mu \\ P &= \Sigma^{-1} \end{aligned} \tag{6}$$

The matrix P is called the *precision matrix*, and its meaning will be explained later.

1.4 Posterior Distribution $P(\theta|D)$

Using Bayes rule, the posterior probability $P(\theta|D)$ can be expressed as

$$P(\theta|D) \propto P(y_{1:t}|x_{1:t},\theta)P(\theta) \propto \left(\prod_{i=1}^{t} P(y_i|x_i,\theta)\right) P(\theta) \tag{7}$$

The y_i's and θ have a diverging relationship at θ, and since θ is unknown, it follows that the y_i's are independent of each other; that is, $P(y_{1:t}|x_{1:t},\theta) = \prod_{i=1}^{t} P(y_i|x_i,\theta)$. We will see that this product can be computed by a simple update rule. First, let's look at the product of $P(y_i|x_i,\theta)P(\theta)$.

$$\begin{aligned}
P(y_i|x_i,\theta)P(\theta) &\propto \exp\{-\frac{1}{2\sigma^2}(y_i-\theta^T x)^2\}\exp\{J^T - \frac{1}{2}\theta^T P\theta\} \\
&\propto \exp\{-\frac{1}{2\sigma^2}(-2y_i\theta^T x_i + \theta^T x_i x_i^T\theta)\}\exp\{J^T\theta - \frac{1}{2}\theta^T P\theta\} \\
&= \exp\{\frac{1}{\sigma^2}y_i x^T\theta - \frac{1}{2\sigma^2}\theta^T x_i {x_i}^T\theta\}\exp\{J^T\theta - \frac{1}{2}\theta^T P\theta\} \\
&= \exp\{(J + \frac{1}{\sigma^2}y_i x_i)^T\theta - \frac{1}{2}\theta^T(P + \frac{1}{\sigma^2}x_i {x_i}^T)\theta\} \\
&= \exp\{J'^T\theta - \frac{1}{2}\theta^T P'\theta\}
\end{aligned}$$

Line 1 to line 2 is true because any term that does not have θ can be absorbed into the normalizer. Now, we can apply the generalized result to (7) and derive

$$P(\theta|D) \propto \exp\{(J + \frac{\sum_i y_i x_i}{\sigma^2})^T\theta - \frac{1}{2}\theta^T(P + \frac{\sum_i x_i {x_i}^T}{\sigma^2})\} \tag{8}$$

So $P(\theta|D)$ is also a normal distribution with $J_{final} = J + \frac{\sum_i y_i x_i}{\sigma^2}$ and $P_{final} = P + \frac{1}{\sigma^2}\sum_i x_i {x_i}^T$. The mean and the covariance of this distribution can be derived with the relation provided earlier:

$$\mu_{final} = \left(\Sigma^{-1} + \frac{\sum_i x_i {x_i}^T}{\sigma^2}\right)^{-1}\frac{\sum_i y_i x_i}{\sigma^2}$$

$$\Sigma_{final} = \left(\Sigma^{-1} + \frac{\sum_i x_i {x_i}^T}{\sigma^2}\right)^{-1}$$

P_{final} is the precision matrix of the normal distribution, and as the number of x_i increases, the terms in this matrix become larger. Also, since P_{final} is the inverse of the covariance, the variance gets lower as the number of samples grow. This is a characteristic of a Gaussian model that a new data point always lowers the variance, but this downgrading of variance does not always make sense. If you believe that there are outliers in your dataset, this model will not work for you.

1.5 Probability Distribution of the Prediction

The next step to compute (3) is to compute $P(y_{t+1}|x_{t+1},\theta)$. Since the linear combination of normal distributions is also a normal distribution, $P(y_{t+1}|x_{t+1},\theta)$ should be in the form $\frac{1}{Z}\exp\{-\frac{1}{2\sigma^2}(y_{t+1} - \mu_{y_{t+1}})^T\Sigma_{y_{t+1}}(y_{t+1}-\mu_{y_{t+1}})\}$, where

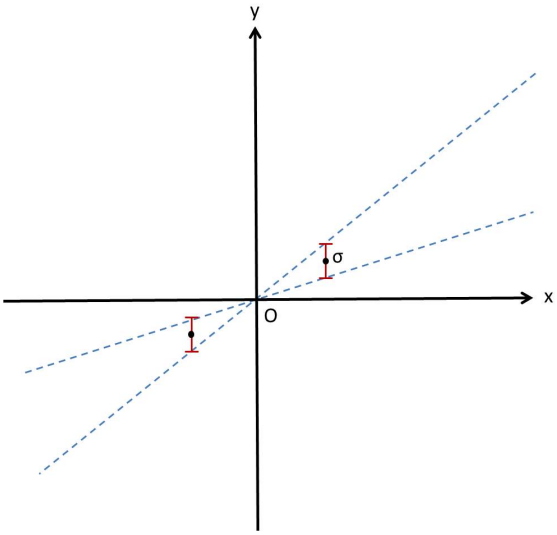
$$\mu_{y_{t+1}} = E[y_{t+1}] = E[\theta^T x_{t+1} + \epsilon] = E[\theta^T x_{t+1}] + E[\epsilon] = E[\theta]^T x_{t+1} + 0 = {\mu_\theta}^T x_{t+1},$$

and

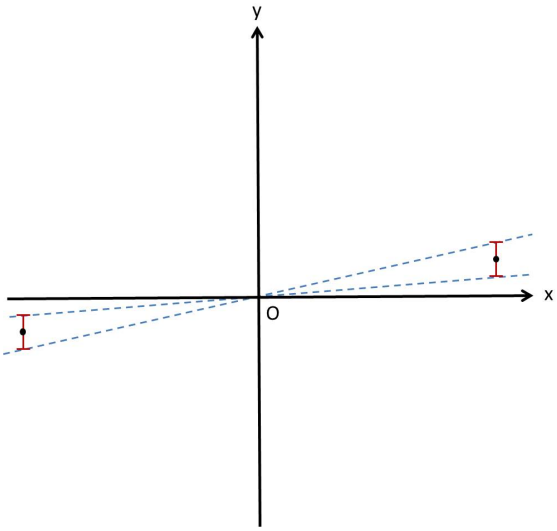
$$\Sigma_{y_{t+1}} = {x_{t+1}}^T \Sigma_\theta x_{t+1} + \sigma^2.$$

The term $x^T \Sigma_\theta x$ measures how large the variance is on the direction that x is on. If x is never observed before, then the variance of the direction of x is large. Also, the variance is not a function of y_{t+1}. The precision is only affected by the input not the output. This is the consequence of having the same σ (observation error) everywhere in the space.

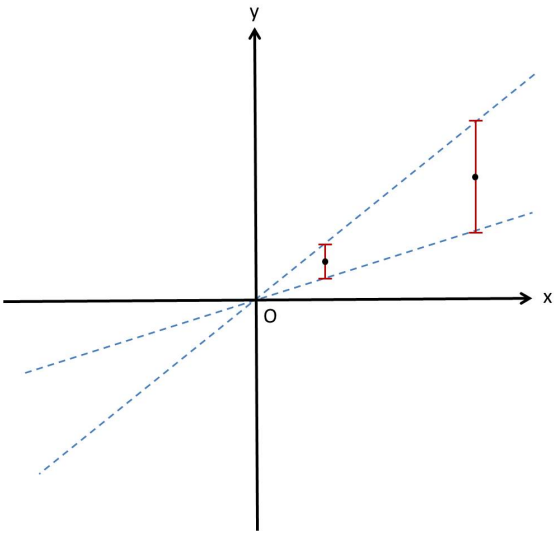
An interesting observation can be made from the expressions of $P(\theta|D)$ and $\Sigma_{y_{t+1}}$. Consider the case in which we are doing linear regression on a set of 2D data (x_i, y_i) and the regression curve must pass the origin. The variance of y_i is σ. When x_i is close to 0, the range of possible values of slope is big, whereas when x_i is large, we are more certain about what the slope can be. But variance of y_i is magnified when the input x_i is large. (See figure 2 for illustration.)



(a) During the training phase, θ has a larger variance when the inputs x_i are small.



(b) Larger inputs decrease the variance of θ.



(c) In testing, θ is know and thus the variance of θ is fixed, and the variance of y rises when the magnitude of x increases.

Figure 2