Bayesian Approach

Jingyuan Sun
ChatGPT Claude

Dec, 2024

1 Bayesian Approach

1.1 Conditional Probability and Bayes' Theorem

1.1.1 Probabilistic Models

A probabilistic model is a set of probability distributions for each point of the sample space. Bayesian methods make this explicit by giving a probability distribution for the model parameters.

1.1.2 Conditional Probability

The conditional probability of an event A given another event B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{where } P(B) > 0.$$

Here:

- $P(A|B)$: The probability of A occurring given that B has occurred.
- $P(A \cap B)$: The probability of both A and B occurring.
- $P(B)$: The probability of B occurring.

A and B are independent if $P(A \cap B) = P(A)P(B)$ *or* $P(A|B) = P(A)$

1.1.3 Bayes' Theorem

Bayes' theorem is a fundamental result in probability theory, which relates the conditional and marginal probabilities of events. It is expressed as:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)},$$

where:

- $P(H|E)$: The posterior probability of hypothesis H given evidence E.
- $P(E|H)$: The likelihood of evidence E given H.

- $P(H)$: The prior probability of H.
- $P(E)$: The marginal probability of E, calculated as:

$$P(E) = \sum_i P(E|H_i)P(H_i),$$

  where $\{H_i\}$ is a partition of the hypothesis space.

1.1.4 Derivation of Bayes' Theorem

Starting from the definition of conditional probability:

$$P(H|E) = \frac{P(H \cap E)}{P(E)} \quad \text{and} \quad P(E|H) = \frac{P(H \cap E)}{P(H)}.$$

Equating $P(H \cap E)$ from both equations:

$$P(H \cap E) = P(H|E)P(E) = P(E|H)P(H)$$

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}.$$

1.1.5 Examples of Bayes' Theorem

Example 1: Disease Testing A disease has a 1% prevalence in the population. A test for this disease has 90% true positive rate and 5% false positive rate. If a person tests positive, what is the probability that they have the disease?

Using Bayes' theorem:

$$P(\text{disease}|\text{positive}) = \frac{P(\text{positive}|\text{disease})P(\text{disease})}{P(\text{positive})},$$

where:

- $P(\text{positive}|\text{disease}) = 0.9$,
- $P(\text{disease}) = 0.01$,
- $P(\text{positive}|\text{no disease}) = 0.05$,
- $P(\text{no disease}) = 0.99$.

The marginal probability of testing positive:

$$P(\text{positive}) = P(\text{positive}|\text{disease})P(\text{disease}) + P(\text{positive}|\text{no disease})P(\text{no disease}),$$

$$P(\text{positive}) = (0.9)(0.01) + (0.05)(0.99) = 0.0594.$$

Therefore:

$$P(\text{disease}|\text{positive}) = \frac{(0.9)(0.01)}{0.0594} \approx 0.15.$$

Example 2: Radioactive Decay A chemist has discovered a new radioactive element and wants to estimate its decay rate. The time between decays ΔT follows an exponential distribution:

$$P(\Delta T) = \lambda e^{-\lambda \Delta T}.$$

The prior probability for λ is also exponential:

$$P(\lambda) = \beta e^{-\beta \lambda}.$$

If a decay interval ΔT_0 is observed, the posterior distribution for λ is:

$$P(\lambda|\Delta T_0) \propto P(\Delta T_0|\lambda)P(\lambda) \propto \lambda e^{-\lambda \Delta T_0} \beta e^{-\beta \lambda}.$$

Combining terms:

$$P(\lambda|\Delta T_0) \propto \lambda e^{-\lambda(\Delta T_0 + \beta)}.$$

This is normalized to a gamma distribution:

$$P(\lambda|\Delta T_0) = \frac{\lambda(\beta + \Delta T_0)^2 e^{-\lambda(\beta + \Delta T_0)}}{\Gamma(2)},$$

where $\Gamma(2) = 1! = 1$.

Gamma Distribution The gamma distribution is defined as:

$$x \sim \text{Gamma}(\alpha, \beta) \implies p(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)},$$

where $\Gamma(\alpha)$ is the gamma function:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

2 Bayesian Inference

2.1 Definition and Explanation

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability of a hypothesis as more evidence or information becomes available.

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

This is a proven result assuming the axioms of probability. Bayesian inference extends this to describe hypotheses, such as:

$$P(hypo|evid) = \frac{P(evid|hypo)P(hypo)}{P(evid|hypo)P(hypo) + P(evid|false\ hypo)P(false\ hypo)}$$

hypo - hypothesis, evid - evidence

Key Points:

- The prior probability P(hypothesis) reflects our belief in the hypothesis before seeing the evidence.
- The likelihood P(evidence—hypothesis) indicates how probable the evidence is, given the hypothesis.
- The posterior probability P(hypothesis—evidence) represents the updated belief after accounting for the evidence.

2.2 Sequential Inference

Bayesian theorem allows for sequential updating of probabilities:

> If new observations are taken, the posterior probability from the first calculation can be used as the prior for the next calculation.

This iterative process ensures that Bayesian inference continuously incorporates new information.

2.3 Conjugate Priors

Conjugate priors simplify Bayesian inference by ensuring that the posterior distribution is of the same family as the prior distribution.

Example: Consider a radioactive element with decay times following an exponential distribution:

$$P(\lambda|\Delta T_0) = \frac{\lambda(\beta + \Delta T_0)^2 e^{-\lambda(\beta+\Delta T_0)}}{\Gamma(2)} \tag{1}$$

If another observation ΔT_1 is made, the posterior distribution becomes:

$$P(\lambda|\Delta T_1) \propto \lambda^2(\beta + \Delta T_0 + \Delta T_1)^3 e^{-\lambda(\beta+\Delta T_0+\Delta T_1)} \tag{2}$$

By normalization, this remains a gamma distribution with updated parameters α' and β'.

Key Insights:

- Conjugate priors maintain computational efficiency.
- The updated posterior always belongs to the gamma family, with parameters directly related to observed data.

2.4 Visualization of Posterior Updates

Below is a visual representation of the sequential updates of the posterior distribution as more observations are incorporated:

3 Markov Chains and Bayesian Networks

3.1 Markov Chains

3.1.1 Definition

A Markov chain is a stochastic model describing a sequence of events where the probability of transitioning to the next state depends only on the current state. This property is called the **Markov property** and is mathematically expressed as:

$$P(X_{n+1}|X_n, X_{n-1}, \ldots, X_1) = P(X_{n+1}|X_n).$$

The dynamics of the chain are described by the *transition matrix* T, where T_{ij} is the probability of transitioning from state i to state j:

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1n} \\ T_{21} & T_{22} & \cdots & T_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{bmatrix}.$$

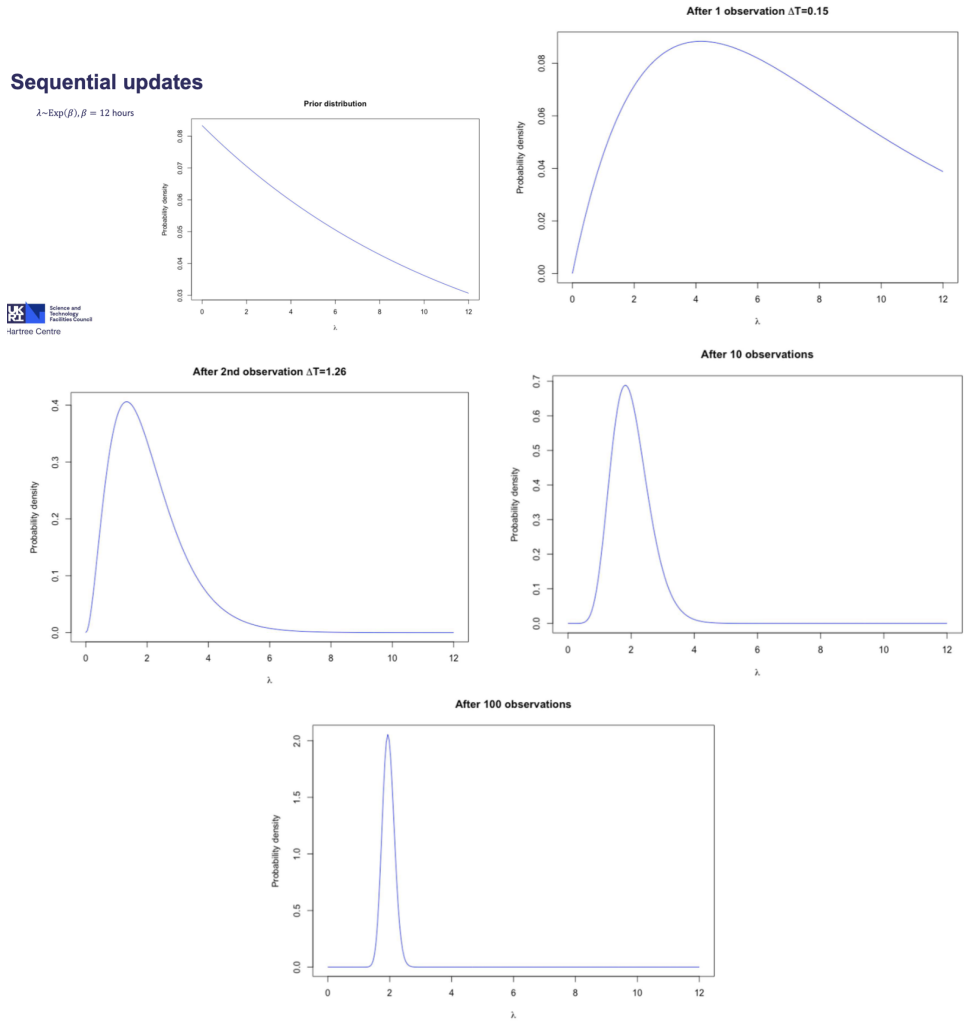

Figure 1: Updates to the posterior distribution after successive observations. The initial prior is progressively refined to reflect new evidence.

3.1.2 Explanation

A Markov chain models processes where the next step depends only on the current position, not the history. This is useful in fields like weather forecasting, where the current state (e.g., sunny or rainy) is sufficient to predict the next state.

3.1.3 Example and Calculation

Consider a Markov chain with four states: A, B, C, D. The transition matrix is:

$$T = \begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix}.$$

If the initial state vector is $\mathbf{p}_0 = [1, 0, 0, 0]$, the distribution after one step is calculated

as:

$$\mathbf{p}_1 = \mathbf{p}_0 T = [1, 0, 0, 0] \begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix} = [0, 0.3, 0.5, 0.2].$$

To compute the distribution after two steps:

$$\mathbf{p}_2 = \mathbf{p}_1 T = [0, 0.3, 0.5, 0.2] \begin{bmatrix} 0 & 0.3 & 0.5 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.1 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix}.$$

Thus, $\mathbf{p}_2 = [0.23, 0.23, 0.42, 0.12]$.

3.2 Conditional Independence

3.2.1 Definition

Two events A and B are conditionally independent given a third event C if:

$$P(A \cap B|C) = P(A|C)P(B|C).$$

anthor way of abtaining:

$$P(A \cap B \cap C)$$

3.2.2 Explanation

Conditional independence implies that given C, the probability of A and B occurring together can be computed as the product of their individual probabilities, reducing the complexity of computation.

3.2.3 Example and Calculation

Consider events A, B, C with:

$$P(C) = 0.5, \quad P(A|C) = 0.3, \quad P(B|C) = 0.4.$$

The joint probability $P(A \cap B|C)$ is:

$$P(A \cap B|C) = P(A|C)P(B|C) = 0.3 \cdot 0.4 = 0.12.$$

Now, calculate the probability of $A \cap B$:

$$P(A \cap B) = P(A \cap B|C)P(C) + P(A \cap B|C^c)P(C^c),$$

where C^c represents the complement of C.

Assume $P(A|C^c) = 0.1$, $P(B|C^c) = 0.2$, and $P(C^c) = 0.5$. Then:

$$P(A \cap B|C^c) = P(A|C^c)P(B|C^c) = 0.1 \cdot 0.2 = 0.02,$$

$$P(A \cap B) = 0.12 \cdot 0.5 + 0.02 \cdot 0.5 = 0.06 + 0.01 = 0.07.$$

3.3 Bayesian Networks

3.3.1 Definition

A Bayesian network is a directed acyclic graph (DAG) representing a joint probability distribution, where:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i)).$$

3.3.2 Explanation

Each node in the graph corresponds to a variable, and edges represent conditional dependencies. The absence of an edge indicates conditional independence between variables.

3.3.3 Example and Calculation

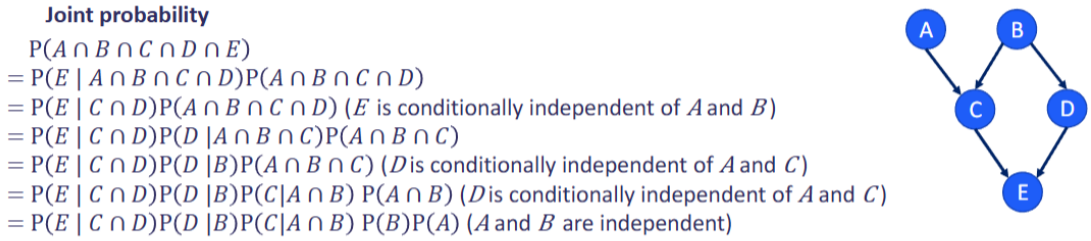

Figure 2: Caption

4 Kalman Filter and Hidden Markov Models

4.1 Kalman Filter

4.1.1 Preliminary Results

Covariance Transformations For a random vector $\mathbf{x}$, its (auto-) covariance matrix is defined as:

$$\mathbf{C} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top.$$

Under a constant matrix transformation $\mathbf{x}' = \mathbf{A}\mathbf{x}$, the covariance transforms as:

$$\mathbf{C}' = \mathbf{A}\mathbf{C}\mathbf{A}^\top.$$

If two random vectors are uncorrelated, their covariance matrix is zero.

Adding Gaussian Distributions If two multivariate normal distributions are multiplied, their product is also a multivariate normal:

$$\mathcal{N}(\mu_0, \Sigma_0) \cdot \mathcal{N}(\mu_1, \Sigma_1) = \mathcal{N}(\mu', \Sigma'),$$

where:

$$\Sigma' = \Sigma_0 - \Sigma_0(\Sigma_0 + \Sigma_1)^{-1}\Sigma_0,$$
$$\mu' = \mu_0 + \Sigma_0(\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0).$$

Alternatively, using $\mathbf{K} = \Sigma_0(\Sigma_0 + \Sigma_1)^{-1}$, we can write:

$$\Sigma' = \Sigma_0 - \mathbf{K}\Sigma_0, \quad \mu' = \mu_0 + \mathbf{K}(\mu_1 - \mu_0).$$

4.1.2 Linear Dynamical Systems

A linear dynamical system evolves according to:

$$\mathbf{x}_{t+1} = \mathbf{F}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t,$$

where $\mathbf{x}$ is the system state and $\mathbf{u}$ is the control input.

Example: Falling Object For a falling tennis ball with position x and velocity v:

$$\begin{bmatrix} x_{t+1} \\ v_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ v_t \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{1}{2}g(\Delta t)^2 \end{bmatrix}.$$

4.1.3 Measurements and Noise

Measurements $\mathbf{y}_t$ are related to the state by:

$$\mathbf{y}_t = \mathbf{M}\mathbf{x}_t + \mathbf{v}_t,$$

where: - $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ (observation noise), - $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ (process noise).

4.1.4 Estimation Process

The Kalman filter follows these steps:

- **Prediction**:
$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}\hat{\mathbf{x}}_{t-1|t-1}, \quad \mathbf{P}_{t|t-1} = \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}^\top + \mathbf{Q}.$$
- **Update**:
$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{M}^\top(\mathbf{M}\mathbf{P}_{t|t-1}\mathbf{M}^\top + \mathbf{R})^{-1},$$
$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{M}\hat{\mathbf{x}}_{t|t-1}),$$
$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{M})\mathbf{P}_{t|t-1}.$$

4.1.5 Algorithm Procedure

The Kalman Filter is a recursive algorithm used to estimate the state of a linear dynamical system from noisy measurements. It consists of four main steps: **Prediction**, **Estimation**, **Observation**, and **Update**. Below, we detail each step along with its mathematical definitions and explanations.

Step 1: Prediction

The prediction step forecasts the state $\mathbf{x}_{t|t-1}$ and the error covariance $\mathbf{P}_{t|t-1}$ for the current time step:

$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{D}\mathbf{u}_t,$$
$$\mathbf{P}_{t|t-1} = \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}^\top + \mathbf{Q}.$$

- $\hat{\mathbf{x}}_{t|t-1}$: Predicted state vector at time t, based on information up to $t-1$.
- $\mathbf{P}_{t|t-1}$: Predicted error covariance matrix, representing the uncertainty in the state prediction.
- $\mathbf{F}$: State transition matrix, modeling the system dynamics.
- $\mathbf{Du}_t$: Control input term, representing external inputs.
- $\mathbf{Q}$: Process noise covariance matrix, accounting for model uncertainty.

Explanation: The system dynamics model is used to predict the current state and uncertainty based on the previous state and control inputs.

—

Step 2: Estimation

In this step, the predicted observation $\mathbf{y}_t$ and its associated uncertainty $\mathbf{S}_t$ are estimated:

$$\hat{\mathbf{y}}_t = \mathbf{M}\hat{\mathbf{x}}_{t|t-1},$$
$$\mathbf{S}_t = \mathbf{M}\mathbf{P}_{t|t-1}\mathbf{M}^\top + \mathbf{R}.$$

- $\hat{\mathbf{y}}_t$: Predicted observation value.
- $\mathbf{S}_t$: Predicted observation covariance matrix.
- $\mathbf{M}$: Observation matrix, mapping the state space to the observation space.
- $\mathbf{R}$: Observation noise covariance matrix, reflecting measurement uncertainty.

Explanation: This step estimates what the observation should look like and its associated uncertainty, providing a basis for comparison with the actual observation.

—

Step 3: Observation

The observation step calculates the residual (also known as the innovation term):

$$\mathbf{v}_t = \mathbf{y}_t - \hat{\mathbf{y}}_t.$$

- $\mathbf{v}_t$: Residual, representing the difference between the actual observation and the predicted observation.

Explanation: The residual captures the deviation between the observed data and the model prediction, forming the basis for updating the state.

—

Step 4: Update

The update step incorporates the residual to refine the state estimate $\hat{\mathbf{x}}_{t|t}$ and the error covariance $\mathbf{P}_{t|t}$:

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{M}^\top\mathbf{S}_t^{-1},$$
$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\mathbf{v}_t,$$
$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{M})\mathbf{P}_{t|t-1}.$$

- $\mathbf{K}_t$: Kalman gain matrix, determining the influence of the residual on the state update.

- $\hat{\mathbf{x}}_{t|t}$: Updated state estimate.
- $\mathbf{P}_{t|t}$: Updated error covariance matrix.

Explanation: The Kalman gain weights the residual to adjust the state estimate, balancing the prediction and the observation uncertainties.

4.2 Hidden Markov Models

4.2.1 Mathematical Definition

HMM is a statistical model where:

- States S_t evolve with transition probabilities $a_{ij} = P(S_t = j|S_{t-1} = i)$,
- Observations O_t are generated with probabilities $b_j(o) = P(O_t = o|S_t = j)$,
- The initial state distribution is $\pi_i = P(S_1 = i)$.

4.2.2 Example and Explanation

Consider weather prediction:

- States: $S_t \in \{\text{Sunny}, \text{Rainy}\}$,
- Observations: $O_t \in \{\text{Dry}, \text{Wet}\}$,
- Transitions and observations are modeled with probabilities.

Forward Algorithm Compute the probability of an observation sequence:

$$\alpha_t(j) = P(O_1, O_2, \ldots, O_t, S_t = j).$$

Viterbi Algorithm Find the most likely sequence of hidden states given the observations.

4.2.3 Markov Chain Monte Carlo (MCMC)

MCMC methods, such as Metropolis-Hastings, sample from the posterior distribution:

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