ECE368: Probabilistic Reasoning

Aman Bhargava

January-April 2021

Contents

0.1 Introduction and Course Information

Course Information

- Professors: Prof. Saeideh Parsaei Fard and Prof. Foad Sohrabi
- Course: Engineering Science, Machine Intelligence Option
- Term: 2021 Winter

Main Course Topics

- Vector, temporal, and spatial models.
- Classification and regression model training.
- Bayesian statistics, frequentist statistics.

Chapter 1

Review Topics

See ECE286 notes for further reference

1.1 Review of Probability Functions

Probability Mass Function: For *discrete random variables*, $P_X(x)$denotes the probability that random variable X takes on value x.

Probability Density Function: For *continuous random variables*, the probability $\Pr\{X \in [x_1, x_2]\}$ is given by $\int_{x_1}^{x_2} f_X(x)dx$.

Joint PMF's and PDF's are similarly defined.

Marginal Probability Distributions: Given joing PMF $P_{X,Y}(x, y)$ or PDF $f_{X,Y}(x, y)$, we can **marginalize** them as follows:

$$P_X(x) = \sum_{y \in Y} P_{X,Y}(x, y) \tag{1.1}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \tag{1.2}$$

Conditional Probability Functions:

$$P_{Y|X}(y, x) = \frac{P_{X,Y}(x, y)}{P_X(x)} \tag{1.3}$$

Prior Probability: Probability **before** an additional observation is made (hence *prior*). Example: $P_X(x)$.

Posterior Probability: Probability **after** an observation is made (hence *post*erior). Example: $P_{X|Y}(x, y)$.

Bayes Rule:

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)} \tag{1.4}$$

1.2 Expectation, Correlation, and Independence

Expectation Value: $\mathbb{E}[x] = \sum_{x \in X} P_X(x) = \int_{-\infty}^{\infty} x f_X(x) dx$

Law or Large Numbers: $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i = \mathbb{E}[X]$

Variance:

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[x])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned} \tag{1.5}$$

Covariance:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}_{XY}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \tag{1.6}$$

Correlation Coefficient:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \tag{1.7}$$

- $\rho_{XY} \in [-1, 1]$
- $\rho > 0$ indicates positive correlation (line of best fit has positive slope).
- $\rho < 0$ indicates negative correlation.
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ **iff** X, Y **are uncorrelated**.

Independence

Theorem 1 *Independence Random variables X, Y are independent **iff***

$$P_{XY}(x, y) = P_X(x) \cdot P_Y(y) \tag{1.8}$$

This also means that $\rho_{XY} = 0$, $P(X|Y) = P(X)$, etc.

1.3 Laws of Large Numbers

Weak Law: Sample mean converges to the mean.

Strong Law: If $\{x_i\}$ are **independent, identically distributed** (i.i.d.) random variables with mean μ, then the **probability of** the sample mean $= \mu$ is 1 as $n \to \infty$.

Chapter 2

Parameter Estimation

2.1 Estimation Terminology

- $\hat{\theta}_n$ is an **estimator** of some unknown parameter θ.
- **Estimation Error:** $\hat{\theta}_n - \theta$
- **Bias** of estimator: $\mathbb{E}[\hat{\theta}_n] - \theta$
  - **Unbiased** estimator: Bias$= 0 = \mathbb{E}[\hat{\theta}_n] - \theta$.
  - **Asymptotically Unbiased:** $\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \theta$ for all θ.
- **Consistency:** Estimator is consistent if $\lim_{n\to\infty} \hat{\theta}_n = \theta$.

2.2 Maximum Likelihood Estimation

Framing: Let random variable $\vec{X} = [X_1, X_2, ..., X_n]$ be defined by either

1. Joint PMF $P_{\vec{X}}(\vec{x};\theta)$
2. Joint PDF $f_{\vec{X}}(\vec{x};\theta)$

$\vec{x}$ is a series of measurements.

Maximum Likelihood Estimation: The ML estimate of model parameter θ is

$$\hat{\theta}_n = \text{argmax}_\theta P_{\vec{X}}(\vec{x};\theta) \tag{2.1}$$

Independent, identically distributed case: If each $x_i \in \vec{x}$ are independent and identically distributed, then

$$P_{\vec{X}}(\vec{x};\theta) = \prod_{i=1}^{n} P_X(x_i;\theta) \tag{2.2}$$

Which we can convert to a summation by taking the **log-likelihood** (recall that logarithm is monotonically increasing, so maximizing log-likelihood is equialent to maximizing likelihood).

$$\hat{\theta}_n = \arg\max_{\theta}(\sum_{i=1}^{n} \log P_X(x_i;\theta)) \tag{2.3}$$

2.3 Frequentist vs. Bayesian Statistics

Frequentist: In **classical statistics**, probability is taken to be approximately equal to the **frequency of events**. Model parameters are assumed to have some deterministic, fixed value (even though they might be unknown).

Bayesian Statistics: Model parameters are treated as **random variables** with their own distributions.

- Generally the more modern approach.
- We are most interested in the **joint probability distribution** of model parameters and model arguments (e.g., $f_x(x,\theta)$).
- **Main criticism:** probabilities are assigned to unrepeatable events (arguably violates the definition of probability as the limit of event frequency).

2.4 Maximum a Posteri Estimation (MAP)

$$\begin{aligned}\hat{\theta}_{map} &= \arg\max_{\theta} f_{\theta|x}(\theta|x) \\ &= \arg\max_{\theta} f_{X|\theta}(x|\theta)\frac{f_\theta(\theta)}{f_X(x)}\end{aligned} \tag{2.4}$$

Where $f_\theta(\theta)$ is the **prior distribution** of model parameter.

- If $f_\theta(\theta)$ is uniform, we will still get the same answer as a **maximum likelihood** estimation.

2.4.1 Picking a Prior Distribution

Best Practice: Pick a distribution of the same form as $f_{X|\theta}(x|\theta)$ (called "conjugate pair").

Beta Distribution: Used for **binomial distribution**.

- Binomial distribution:

$$P_{X=k|\theta} = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{2.5}$$

  where

  - θ: Probability of success on each Bernoulli trial.
  - n: Total number of trials.
  - k: Total number of successful trials.

- **Beta Distribution:**

$$f_\theta(\theta; \alpha, \beta) = \begin{cases} c\theta^{\alpha-1}(1-\theta)^{\beta-1} & \text{for } \theta \in [0,1] \\ 0 & \text{else} \end{cases} \tag{2.6}$$

  Where

  - α, β are customizable parameters.
  - $c = [\Gamma(\alpha+\beta)]/[\Gamma(\alpha)\Gamma(\beta)]$
  - $\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du$
  - $\Gamma(x+1) = x\Gamma(x)$ for all $x \in \mathbb{R}$.
  - $\Gamma(n+1) = n!$ for integer n.
  - $\therefore c = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$ for integer α, β.
  - $\mu_f = \mathbb{E}[f_\theta(\theta)] = \frac{\alpha}{\alpha+\beta}$
  - Maximum likelihood $\arg\max_\theta f_\theta(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$

2.5 Conditional Expectation Estimator

Key Idea: Find the **expected value** for the estimator given your observations.

$$\hat{\theta}_{conditionalexpectation} = \mathbb{E}[\theta|\vec{X} = \vec{x}] = \int_{-\infty}^{\infty} \theta f_{\theta|\vec{x}}(\theta|\vec{x}) \tag{2.7}$$

2.6 Bayesian Least Mean Square Estimator (LMS)

Key Idea: To estimate random variable model parameter θ, we find

$$\hat{\theta}_{LMS} = \arg\min_{\hat{\theta}} \mathbb{E}[(\theta - \hat{\theta})^2] \tag{2.8}$$

- $\hat{\theta}_{LMS} = \mathbb{E}[\theta]$ achieves the goal.
- **Equivalently:** We can also find

$$\hat{\theta}_{LMS} = \arg\min_{\hat{\theta}} (\mathbb{E}[\theta - \hat{\theta}])^2 \tag{2.9}$$

2.7 LMS with Observations

...