EECS 445: Machine Learning **Fall 2016**

Discussion 2: Probability & Bayesian Paradigm

Written by: Benjamin Bray and Chansoo Lee

2.1 Writing Mathematical Proofs

Online Resources:

- Prof. John Lee at University of Washington has an excellent writeup focusing on writing techniques. [link]
- Prof. Michael Hutchings at UC Berkeley has a more beginner-friendly writeup which includes basic proof templates. [link]

Important things

- Do NOT start by stating the conclusion of an implication you are trying to prove. When you prove p implies q, q should NEVER appear until the very end of the proof.
- Define variables (e.g. let a be a real number) when they appear first time. In C/JAVA, you'd get a compile time error which prevents you from even *submitting* a proof with undefined variables. In math proofs, you will get instead a *run-time* error when readers try to understand your proof, like in Python. Remember that variables have no meaning until you define them.
- Explain your logic with English sentences instead of ambiguous math symbols such as $\therefore$ and $\because$. You wouldn't want your collaborators to write even more code when they *document* their code.

2.2 Probability Notations

We will review the essential probability definitions using two-dice example.

- **Random variable** is denoted by capital letters such as X and Y.
- *Sample space* $\Omega(X)$ is the *set* of values that X can take.
- An **Event** is a subset of sample space. Events are also denoted by capital letters such as A or B. We often omit the set notations. For example, $(X = x)$ is the event $\{x\}$, and $(a \leq X \leq b)$ is the event $\{x \in \Omega(X) : a \leq x \leq b\}$.
- **Probability of an event** is denoted by $P(A)$.
- **Joint probability**, denoted by $P(A, B)$, is the probability that the event A and B both happen.
- **Conditional probability** of an event, $P(A|B) = P(A \cap B)/P(B)$ is the probability that A happens given B happens.
- Two events are **independent** if $P(A|B) = P(A)$. *Question:* What does it imply about $P(A \cap B)$?
- **Probability mass function (pmf)** of a discrete random variable, denoted by f_X, μ_X, or p_X, maps each element $x \in \Omega(X)$ to the probability that $X = x$.

- **Joint probability mass function** of two discrete random variables, denoted by $f_{X,Y}(x, y)$, maps each element in $(x, y) \in \Omega(X) \times \Omega(Y)$ to the probability that $(X, Y) = (x, y)$.
- **Conditional mass function** is defined as $f_{X|Y}(x|y) = f_{X,Y}(x, y)/f_Y(y)$.
- **Expectation** of X is defined as:
$$\mathbb{E}[X] = \sum_{x \in \Omega(X)} f_X(x)x$$
- **Conditional expectation** of X given Y is a function of Y defined as:
$$\mathbb{E}[X|Y] = \sum_{x \in \Omega(X)} f_{X|Y}(x)x$$
- **Variance** of X is defined as: $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}(X))^2]$.

Excercise 2.1. Throw two dice and define X_1 and X_2 to be the result of each. Let $M = \max(X_1, X_2)$ and $S = X_1 + X_2$. Find

- $\mathbb{E}[X_2|X_1]$
- $\mathbb{E}[S|X_1]$
- $\mathbb{E}[M|X_1]$
- $\mathbb{E}[S|X_2 \text{ is even}]$

Excercise 2.2. Let X, Y, Z be random variables and a, b be real numbers. Show the linearly of conditional expectation: $\mathbb{E}[aY + bZ|X] = a\mathbb{E}[Y|X] + b\mathbb{E}[Z|X]$.

Excercise 2.3. Show that $\text{Var}(X|Y) = E(X^2|Y) - [E(X|Y)]^2$.

Remarks

- We sometimes omit the subscripts for pmfs and deduce them from the context. For example, $f(x, y)$ is the joint density function.
- For a continuous random variable, we have a **(conditional/joint) probability density function** instead. But we can't really explain it in this course. Generally speaking, however, things that hold for pmfs also hold for pdfs, after converting summations $\sum_{x \in \Omega(X)}$ to integrals $\int_{\Omega(X)} dx$.
- One important difference is that the probability of event $X = x$ for any $x \in \Omega(X)$ is 0, so pdf does *not* define the probability that $X = x$.

For more on discrete random variables, read this. For more on continuous random variables, read this.

2.3 Two Approaches to Data Analysis

There are two cultures in the use of statistical modeling to reach conclusions from data [Breiman, 2001], each differing in their goals and assumptions. We are typically given a vector of input variables $\mathcal{X}$ and the corresponding response variables $\mathcal{Y}$, which have been created by nature in response to the input variables. Once we have a dataset, we may be interested in either the **prediction** of response variables for future input variables or in extracting **information** about how nature is associating the response variables with the input variables from within its black box.

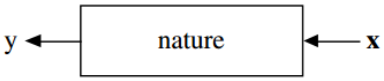


Figure 2.1: Data is generated by a black box that we cannot peek inside of.

Data Modeling Culture

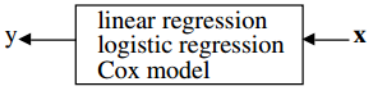
There are two different approaches toward these goals. Analysis in the **data modeling culture** (i.e. most statisticians) makes assumptions about the data is generated from within nature's black box, often using a statistical model to do so. For example, a common data model is that observed data is generated by independent draws from a known family of probability distributions, with some parameters θ that can be tuned to fit the data.

$$\begin{aligned} \mathcal{X} &\quad \text{input variables} \\ \varepsilon &\quad \text{random noise} \\ \theta &\quad \text{parameters} \\ \mathcal{Y} &\sim f(\mathcal{X}, \varepsilon, \theta) \end{aligned}$$

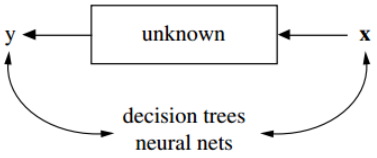
Once the values of the parameters are estimated from the data, the model can be used for information or prediction. Different data models work best in different settings, and care must be taken to choose an appropriate model. Statisticians use *goodness-of-fit* tests to check that their data models do not make wildly inappropriate assumptions about the data. Many classical statistical models fall under this category, including linear regression, logistic regression, mixture models, and generative probabilistic graphical models in general.

Algorithmic Modeling Culture

Analysis in the **algorithmic modeling culture** (i.e. most computer scientists) assumes that the inside of the black box is far too complex and unknown to model directly. Instead, their approach is to find an algorithm $f(\mathcal{X})$ that operates on the input variables to predict the response variables. Because algorithmic modelers are willing to sacrifice interpretability of their models for higher accuracy, they are more able to take advantage of powerful, one-size-fits-most techniques like deep neural nets and random forest models.



(a) Data modelers make assumptions about how the black box generates the response variables. If the model does not reflect nature, this approach will perform poorly.



(b) Algorithmic modelers don't care whether or not their model reflects the data generating process, only that it performs well at prediction tasks.

Which Approach is Better?

Each approach to data analysis has its merits, and this is a vastly oversimplified view of how researchers and data scientists solve problems. Data models give statisticians a peek into nature's black box, and allow

for greater interpretability. It is also much easier to provide confidence intervals for a data model, which can be important in domains like medicine. Algorithmic models are more capable of modeling complicated data like images and speech, although it is harder to make guarantees about the results. This semester, we will give you a chance to solve problems using both approaches, and you can weigh for yourself the relative advantages and disadvantages of each.

2.4 Probabilistic Modeling: Maximum Likelihood

Let $\mathcal{D}$ be a family of distributions over n samples, parametrized by θ, where each element has joint pmf(pdf) $p(\cdot;\theta)$ over n samples. **Likelihood** is a function of the distribution parameter θ given samples (data) $X_1, \ldots, X_n$:

$$L(\theta; X_1, \ldots, X_n) = p(X_i, \ldots, X_n; \theta).$$

We mostly consider a special case of *product distributions*, where $\mu_\theta(X_i, \ldots, X_n) = \prod_{i=1}^{N} p(X_i;\theta)$.

The maximum likelihood estimate $\hat{\theta}_{ML}$ is simply the maximizer for $L(\theta)$. It is the parameter value under which the observed data is most probable.

Discuss: Instead of the likelihood, we use the log likelihood $\mathcal{L}(\theta|\mathcal{X}) = \log P(\theta|\mathcal{X})$. Can you explain why?

Excercise 2.4. (Discrete) Suppose we observe the result of N flips of a coin with unknown bias θ (probability of heads).

1. Write the likelihood function for this example.
2. Differentiate to find the MLE.

Remark When asked to estimate parameters, it always helps to write down the corresponding data model. This involves identifying the relevant probability distributions and any other variables that may influence the result. Here, we can use the **Bernoulli distribution** Ber$[\theta]$ to model the result of a single coin flip.

$$\theta \sim \text{fixed parameter} \qquad \text{(bias of coin)}$$
$$X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Ber}[\theta] \qquad \text{(coin flips)}$$

Excercise 2.5. (Continuous and bounded) Suppose we observe N samples from uniform distribution over $[0, \theta]$. What is the MLE of θ? Hint: $L(\theta) = \theta^{-n}$.

Excercise 2.6. (Continuous and unbounded) Suppose we observe N samples from the density function $p(x;\theta) = \frac{1}{2\theta}\exp(-|x|/\theta)$. What is the MLE of θ?

Remark Remember that θ can be any mathematical object, such as:

- scalar, as in all examples above.
- vector, e.g. $(\mu, \sigma) \in \mathbb{R}^2$ when $\mathcal{D}$ is one-dimensional Gaussian.
- vector and matrix, e.g. $(\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{R}^{m\times m}$ for mean vector and covariance matrix when $\mathcal{D}$ is multidimensional Gaussian.

If L is a function of more than one variable, then the MLE requires optimizing over all variables.

References

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16(3):199–231, 08 2001. doi: 10.1214/ss/1009213726. URL http://dx.doi.org/10.1214/ss/1009213726.