CS578 Statistical Machine Learning Lecture 11

Jean Honorio
Purdue University

Today's topics

- Probability review
  - joint probability
  - marginal probability
  - conditional probability
- Independence
- Maximum likelihood estimation

Joint probability

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables

  e.g., P(warning, weather) = a 2 × 4 matrix of values:

	weather = sunny	weather = rainy	weather = cloudy	weather= snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Marginal probability

- **Marginal** (or unconditional) probability corresponds to belief that event will occur regardless of conditioning events

- Marginalization: $$P(A) = \sum_b P(A, B = b)$$

- Example: What is P(weather=cloudy)?

	weather = sunny	weather = rainy	weather = cloudy	weather= snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

- P(weather=cloudy)
  = P(weather=cloudy, warning=Y) + P(weather=cloudy, warning=N)
  = 0.02 + 0.31 = 0.33

Conditional probability

- **Conditional** (or posterior) probability:
  - e.g., P(warning=Y | weather=snow) = 0.4
  - Complete conditional distributions specify conditional probability for all possible combinations of a set of RVs:

    P(warning | weather) =
    {P(warning=Y | weather=sunny), P(warning=N | weather=sunny),
    P(warning=Y | weather=rainy), P(warning=N | weather=rainy),
    P(warning=Y | weather=cloudy), P(warning=N | weather=cloudy),
    P(warning=Y | weather=snow), P(warning=N | weather=snow)}

Conditional probability

- Definition of conditional probability:

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

- **Product rule** gives an alternative formulation:

$$\begin{aligned} P(A,B) &= P(A \mid B)P(B) \\ &= P(B \mid A)P(A) \end{aligned}$$

- **Bayes rule** uses the product rule:

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

Example

- Conditional probability:

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

- Example: What is P(weather = sunny | warning = Y)?

	weather = sunny	weather = rainy	weather = cloudy	weather= snow
$\text{warning} = Y$	0.005	0.08	0.02	0.02
$\text{warning} = N$	0.415	0.12	0.31	0.03

- P(warning=Y) = 0.005 + 0.08 + 0.02 + 0.02 = 0.125 (marginal probability)

- P(weather=sunny | warning=Y)

  = P(weather=sunny, warning=Y) / P(warning=Y)

  = 0.005 / 0.125 = 0.04

Conditional probability

- **Chain rule** is derived by successive application of product rule:

$$
\begin{aligned}
P(X_1,\ldots,X_n) &= P(X_n|X_1,\ldots,X_{n-1})P(X_1,\ldots,X_{n-1}) \\
&= P(X_n|X_1,\ldots,X_{n-1})P(X_{n-1}|X_1,\ldots,X_{n-2})P(X_1,\ldots,X_{n-2}) \\
&= \ldots \\
&= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1})
\end{aligned}
$$

Today's topics

- Probability review
  - joint probability
  - marginal probability
  - conditional probability
- Independence
- Maximum likelihood estimation

Recall that in general…

- Joint probability $$P(A,B)$$
- Marginal probability $$P(A) = \sum_b P(A, B = b)$$
- Conditional probability $$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

$$P(A,B) = P(A \mid B)P(B)$$

Independence

- A and B are independent iff for all values of *A, B*:
  - $P(A, B) = P(A)\,P(B)$
  - Equivalently: $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$
  - *Knowing B tells you nothing about A*
- Examples
  - Coin flip 1 and coin flip 2
  - Weather and coin flip

Example of independent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.15	0.075	→ P(X=1) = 0.25
X = 2	0.075	0.45	0.225	→ P(X=2) = 0.75
	↓ P(Y=1) = 0.1	↓ P(Y=2) = 0.6	↓ P(Y=3) = 0.3	

- P(X=1,Y=1) = P(X=1) P(Y=1) ?   P(X=2,Y=1) = P(X=2) P(Y=1) ?
- P(X=1,Y=2) = P(X=1) P(Y=2) ?   P(X=2,Y=2) = P(X=2) P(Y=2) ?
- P(X=1,Y=3) = P(X=1) P(Y=3) ?   P(X=2,Y=3) = P(X=2) P(Y=3) ?

- If the answer to the 6 questions is “Yes”, then X and Y are independent

Example of independent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.15	0.075	→ P(X=1) = 0.25
X = 2	0.075	0.45	0.225	→ P(X=2) = 0.75
	↓ P(Y=1) = 0.1	↓ P(Y=2) = 0.6	↓ P(Y=3) = 0.3	

- 0.025 = 0.25 * 0.1 (Yes)
- 0.15 = 0.25 * 0.6 (Yes)
- 0.075 = 0.25 * 0.3 (Yes)

0.075 = 0.75 * 0.1 (Yes)

0.45 = 0.75 * 0.6 (Yes)

0.225 = 0.75 * 0.3 (Yes)

- The answer to the 6 questions is "Yes". X and Y are independent.

Example of dependent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.125	0.1	$\rightarrow$ P(X=1) = 0.25
X = 2	0.075	0.475	0.2	$\rightarrow$ P(X=2) = 0.75
	$\downarrow$ P(Y=1) = 0.1	$\downarrow$ P(Y=2) = 0.6	$\downarrow$ P(Y=3) = 0.3	

- P(X=1,Y=1) = P(X=1) P(Y=1) ?　　P(X=2,Y=1) = P(X=2) P(Y=1) ?
- P(X=1,Y=2) = P(X=1) P(Y=2) ?　　P(X=2,Y=2) = P(X=2) P(Y=2) ?
- P(X=1,Y=3) = P(X=1) P(Y=3) ?　　P(X=2,Y=3) = P(X=2) P(Y=3) ?

- If the answer to the 6 questions is “Yes”, then X and Y are independent

Example of dependent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.125	0.1	P(X=1) = 0.25
X = 2	0.075	0.475	0.2	P(X=2) = 0.75
	P(Y=1) = 0.1	P(Y=2) = 0.6	P(Y=3) = 0.3	

- 0.025 = 0.25 * 0.1 (Yes) 0.075 = 0.75 * 0.1 (Yes)
- 0.125 = 0.25 * 0.6 (No) 0.475 = 0.75 * 0.6 (No)
- 0.1 = 0.25 * 0.3 (No) 0.2 = 0.75 * 0.3 (No)
- The answer to at least 1 question is "No". X and Y are NOT independent.

Conditional independence

- Two variables *A* and *B* are **conditionally** independent given *Z* iff for all values of *A, B, Z:*

    - $P(A, B \mid Z) = P(A \mid Z) P(B \mid Z)$

    - Equivalently: $P(A \mid B, Z) = P(A \mid Z)$ or $P(B \mid A, Z) = P(B \mid Z)$

- *Note: independence does not imply conditional independence or vice versa*

Today's topics

- Probability review
  - joint probability
  - marginal probability
  - conditional probability
- Independence
- Maximum likelihood estimation

Likelihood function

- A random variable $\underline{x}$ has **parameters** θ and probability $P(\underline{x};\theta)$

  e.g., Bernoulli: $\theta = p \,, \quad P(x;\theta) = p^x (1-p)^{1-x}$

  Normal: $\theta = (\underline{\mu}, \Sigma) \,, \quad P(\underline{x};\theta) = (2\pi)^{-d/2} \left|\Sigma\right|^{-1/2} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})}$

- Assume we have *n* **independent** samples $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$

- Define the dataset $D = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$

- The likelihood function represents the probability of the dataset *D* as a function of the model parameters θ

$$L(D;\theta) = P(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n;\theta) \underset{\text{by independence}}{=} \prod_{i=1}^{n} P(\underline{x}_i;\theta)$$

Likelihood function

- The likelihood function represents the probability of the dataset *D* as a function of the model parameters θ

$$L(D;\theta) = P(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n; \theta) = \prod_{i=1}^{n} P(\underline{x}_i; \theta)$$

- Gives relative probability of data given a parameter
- We can compare two models θ and θ' by comparing their likelihoods
- We say that model θ is better for explaining the dataset *D* than model θ' if

$$L(D;\theta) > L(D;\theta')$$

Maximum likelihood estimation (MLE)

- Most widely used method of parameter estimation
- **Intuition:** a model with higher likelihood explains better the data
- "Learn" the best parameters θ that maximizes likelihood:

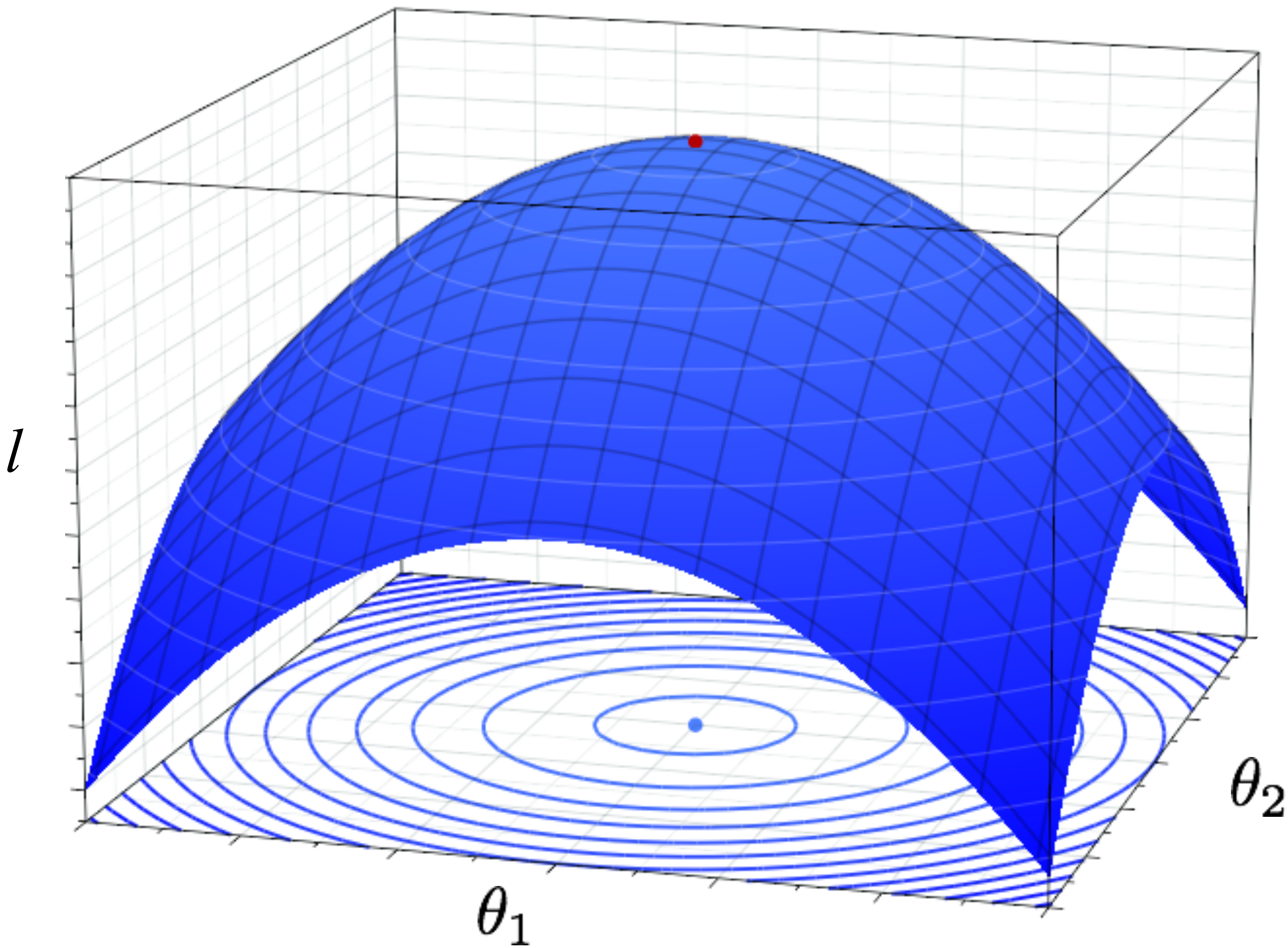
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L(D;\theta)$$

- Often easier to work with log-likelihood:

$$l(D;\theta) = \log L(D;\theta) = \log \prod_{i=1}^{n} P(\underline{x}_i;\theta) = \sum_{i=1}^{n} \log P(\underline{x}_i;\theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ l(D;\theta)$$

Likelihood surface



If the log-likelihood surface is concave, we can often determine the parameters that maximize the function analytically

Maximum Likelihood Estimation (MLE) for Bernoulli

- For a Bernoulli r.v. $x_i \in \{0,1\}$, $\theta = p$, $P(x_i;\theta) = p^{x_i}(1-p)^{1-x_i}$
- Clearly: $\log P(x_i;\theta) = x_i \log p + (1-x_i)\log(1-p)$
- The **log-likelihood function** is:

$$l(D;\theta) = \sum_{i=1}^{n} \log P(\underline{x}_i;\theta)$$

$$= \sum_{i=1}^{n} \left(x_i \log p + (1-x_i)\log(1-p) \right)$$

$$= \left(\sum_{i=1}^{n} x_i \right) \log p + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-p)$$

- Recall that the **MLE** is:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ l(D;\theta)$$

Maximum Likelihood Estimation (MLE) for Bernoulli

- For a Bernoulli r.v. $x_i \in \{0,1\}$, $\theta = p$, $P(x_i;\theta) = p^{x_i}(1-p)^{1-x_i}$

- The **log-likelihood function** is:

$$l(D;\theta) = \left(\sum_{i=1}^{n} x_i\right)\log p + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p)$$

- Recall that the **MLE** is: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ l(D;\theta)$

- We can maximize $l(D;\theta)$ by taking derivative equal to zero:

$$\frac{\partial l(D;\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \qquad \text{then} \qquad \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The MLE $\hat{\theta} = \hat{p}$ is the proportion of ones in the dataset. This is intuitive since the parameter $\theta = p = \mathrm{E}[X]$ is the expected proportion of ones.

Maximum Likelihood Estimation (MLE) for Bernoulli

```
>> example_bernoulli = @(n) mean(rand(1,n)>0.5);

>> example_bernoulli(10)

ans =
   0.7

>> example_bernoulli(100)

ans =
   0.53

>> example_bernoulli(10000)

ans =
   0.5052
```