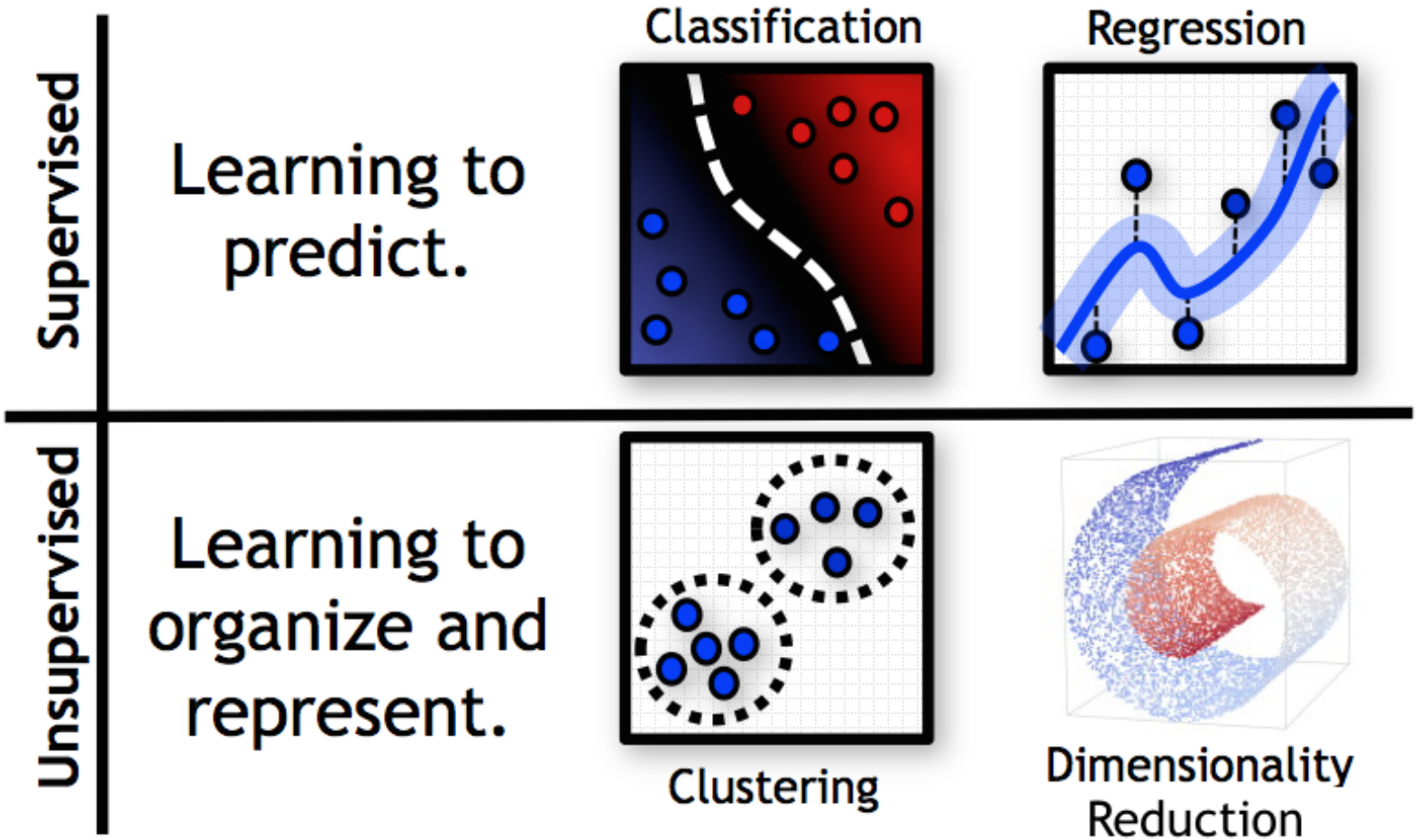
COMPSCI 689
Lecture 17: Joint Probability Models

Benjamin M. Marlin

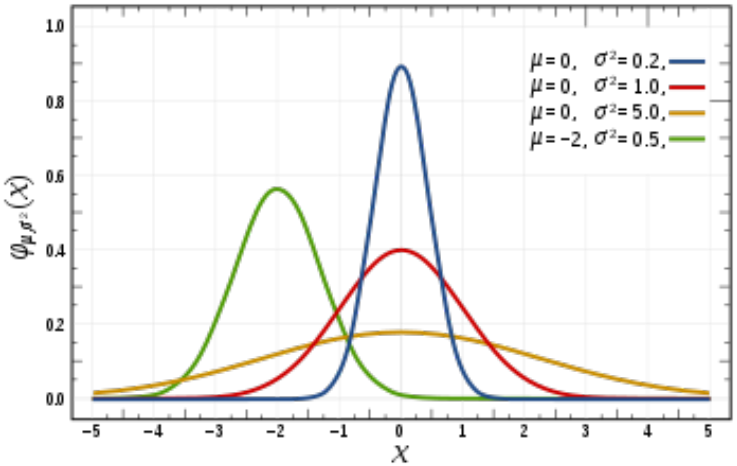
College of Information and Computer Sciences
University of Massachusetts Amherst

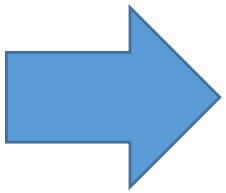
Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

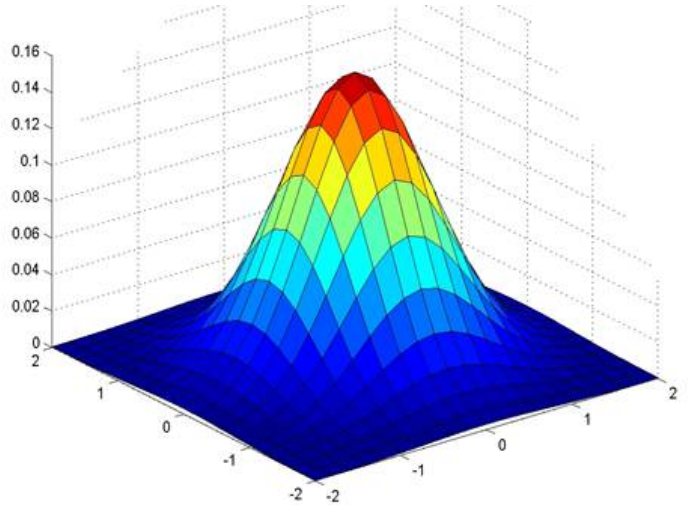
Machine Learning Tasks

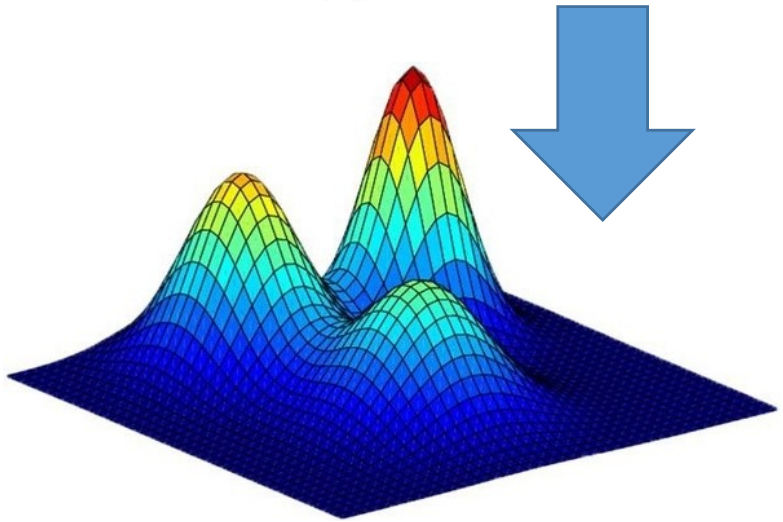


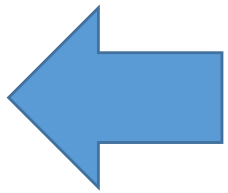
Probabilistic Unsupervised Learning

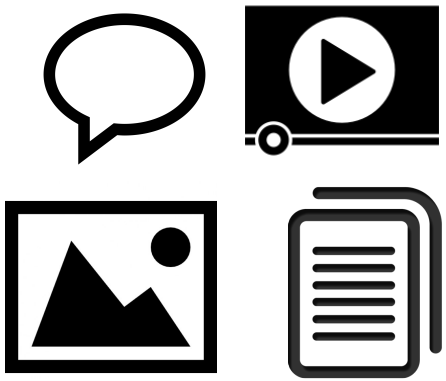












Vector-Valued Random Variables

- In probabilistic unsupervised learning, our goal is to model multivariate data $\mathbf{x} = [x_1, ..., x_D]$ generated by an unknown process using a probabilistic model learned from a data set $\mathcal{D} = \{\mathbf{x}_n | 1 \leq n \leq N\}$.
- Since the data are vectors, we use vector-valued random variables to model them $\mathbf{X} = [X_1, ..., X_D]$.
- Each data dimension d takes values from a potentially different set $\mathcal{X}_d$. We have $\mathbf{x} \in \mathcal{X}$. $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_D$.

Discrete Joint Distributions

- A probability distribution over the joint settings of multiple random variables (or equivalently a vector-valued random variable) is referred to as a *joint distribution*.
- When all dimensions of $\mathbf{x}$ are discrete, the joint distribution is represented by a *joint probability mass* function $P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, ..., .X_D = x_d)$.

Example: Discrete Joint Distribution

x	P(X=x|θ)
[0,0,0,0,0]	θ_0
[0,0,0,0,1]	θ_1
[0,0,0,1,0]	θ_2
[0,0,0,1,1]	θ_3
⋮	
[1,1,1,1,1]	θ_{31}

Discrete Joint Distributions

- Joint probability mass functions need to satisfy $P(\mathbf{X} = \mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{X}$ and $\sum_{\mathbf{x} \in \mathcal{X}} P(\mathbf{X} = \mathbf{x}) = 1$.
- The normalization expression unpacks to:

$$\sum_{\mathbf{x} \in \mathcal{X}} P(\mathbf{X} = \mathbf{x}) = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_D \in \mathcal{X}_D} P(X_1 = x_1, ..., X_D = x_D) = 1$$

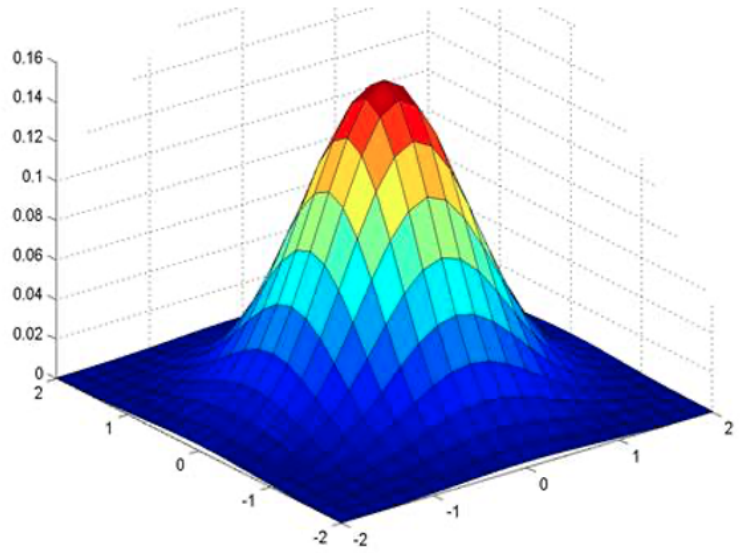
- In the case of a joint distribution over D binary variables, this summation is over all 2^D joint configurations of the D binary variables.

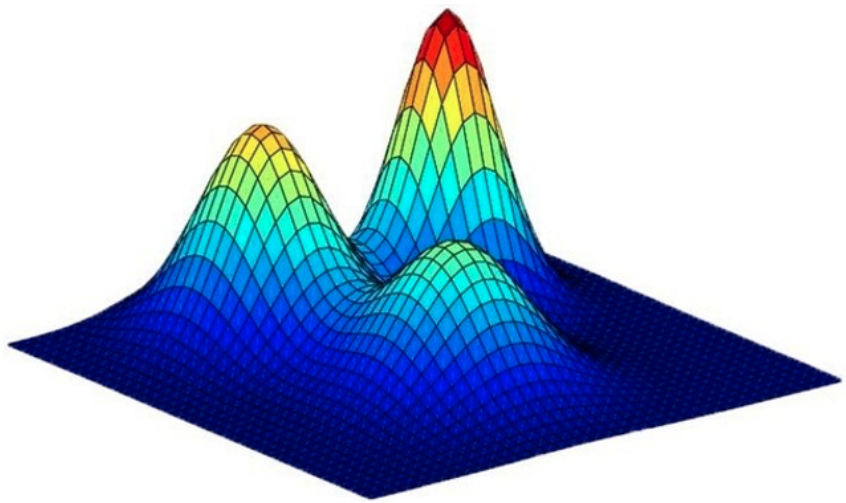
Continuous Joint Distributions

- When all dimensions of $\mathbf{x}$ are continuous, the joint distribution is represented by a *joint probability density* function $p(\mathbf{X} = \mathbf{x}) = p(X_1 = x_1, ..., .X_D = x_d)$.
- Joint probability density functions need to satisfy $p(\mathbf{X} = \mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{X}$ and $\int_{\mathcal{X}} p(\mathbf{X} = \mathbf{x}) d\mathbf{x} = 1$.
- The normalization expression unpacks to:

$$\int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_D} p(X_1 = x_1, ..., X_D = x_D) dx_1 \cdots dx_D = 1$$

Example: Continuous Joint Distribution





Mixed Joint Distributions

- When the data are of mixed-type, we can still model them via a probability distribution consisting of both mass and density function components.
- We'll come back to this later in the lecture.

Probabilistic Inference

- Joint probability distributions allow us to compute the joint probability (mass or density) of a fully specified vector of random variables $\mathbf{x} = [x_1, ..., x_D]$.
- Often, we are instead interested in computing the probability of a subset of the random variables, an operation referred to as *marginalization*.
- We can also use joint distributions to make probabilistic predictions about the distribution of one subset of random variables in the joint distribution given values for another subset. This operation is called *conditioning*.
- Marginalization and conditioning are the two fundamental probabilistic inference operations.

Marginalization: Discrete Case

- Suppose we have a joint probability mass function on a vector-valued random variable $\mathbf{X}$. Let $A \subseteq \{1, ..., D\}$, $M = |A|$, and $\mathbf{X}_A = [X_{A_1}, ..., X_{A_M}]$.
- The probability mass function of $P(\mathbf{X}_A = \mathbf{x}_A)$ specifies the *marginal distribution* of $\mathbf{X}_A$.
- Let $B = \{1, ..., D\} \backslash A$. The marginal distribution of $\mathbf{X}_A$ is then given by:

$$P(\mathbf{X}_A = \mathbf{x}_A) = \sum_{\mathbf{x}_B \in \mathcal{X}_B} P(\mathbf{X}_A = \mathbf{x}_A, \mathbf{X}_B = \mathbf{x}_B)$$

Example: Discrete Marginalization

$\mathbf{x}=[x_1,x_2,x_3]$	$P(\mathbf{X}=\mathbf{x}\mid\theta)$
[0,0,0]	θ_0
[0,0,1]	θ_1
[0,1,0]	θ_2
[0,1,1]	θ_3
[1,0,0]	θ_4
[1,0,1]	θ_5
[1,1,0]	θ_6
[1,1,1]	θ_7

$$
\begin{aligned}
P(X_1=1\mid\theta) &= P(\mathbf{X}=[1,0,0]\mid\theta) \\
&\quad + P(\mathbf{X}=[1,0,1]\mid\theta) \\
&\quad + P(\mathbf{X}=[1,1,0]\mid\theta) \\
&\quad + P(\mathbf{X}=[1,1,1]\mid\theta) \\
&= \theta_4 + \theta_5 + \theta_6 + \theta_7
\end{aligned}
$$

Conditioning: Discrete Case

- Suppose we have a joint probability mass function on a vector-valued random variable $\mathbf{X} \in \mathbb{R}^D$. Let $A \subseteq \{1, ..., D\}$ and let $B = \{1, ..., D\} \backslash A$.
- The *conditional probability mass function* of $\mathbf{X}_A$ given $\mathbf{X}_B$ is defined as shown below:

$$P(\mathbf{X}_A = \mathbf{x}_A | \mathbf{X}_B = \mathbf{x}_B) = \frac{P(\mathbf{X}_A = \mathbf{x}_A, \mathbf{X}_B = \mathbf{x}_B)}{P(\mathbf{X}_B = \mathbf{x}_B)}$$

- This definition follows from the definition of conditional probability for events.
- Note that the numerator is the joint mass function and the denominator is the marginal mass function of $\mathbf{X}_B$.

Example: Discrete Conditioning

$\mathbf{x}=[x_1,x_2,x_3]$	$\mathbf{P(X=x\mid\theta)}$
[0,0,0]	θ_0
[0,0,1]	θ_1
[0,1,0]	θ_2
[0,1,1]	θ_3
[1,0,0]	θ_4
[1,0,1]	θ_5
[1,1,0]	θ_6
[1,1,1]	θ_7

$$P(X_2=0, X_3=0 \mid X_1=1, \theta)$$
$$= P(\mathbf{X}=[1,0,0]\mid\theta) / P(X_1=1\mid\theta)$$
$$= \theta_4 / (\theta_4 + \theta_5 + \theta_6 + \theta_7)$$

Marginalization: Continuous Case

- Suppose we have a joint probability density function on a vector-valued random variable $\mathbf{X} \in \mathbb{R}^D$. Let $A \subseteq \{1, ..., D\}$, $M = |A|$, and $\mathbf{X}_A = [X_{A_1}, ..., X_{A_M}]$.
- The probability density function $p(\mathbf{X}_A = \mathbf{x}_A)$ specifies the *marginal distribution* of $\mathbf{X}_A$.
- Let $B = \{1, ..., D\} \backslash A$. The marginal distribution of $\mathbf{X}_A$ is then given by:

$$p(\mathbf{X}_A = \mathbf{x}_A) = \int_{\mathcal{X}_B} p(\mathbf{X}_A = \mathbf{x}_A, \mathbf{X}_B = \mathbf{x}_B) d\mathbf{x}_B$$

$p(X_A = x_A) \int_{X_B} (X_A = x_A | Z = 2, X_B = x_B | Z = 1) dX_B$

Conditioning: Continuous Case

- Suppose we have a joint probability density function on a vector-valued random variable $\mathbf{X} \in \mathbb{R}^D$. Let $A \subseteq \{1, ..., D\}$ and let $B = \{1, ..., D\} \backslash A$.
- The *conditional probability density* of $\mathbf{X}_A$ given $\mathbf{X}_B$ is defined as shown below:

$$p(\mathbf{X}_A = \mathbf{x}_A | \mathbf{X}_B = \mathbf{x}_B) = \frac{p(\mathbf{X}_A = \mathbf{x}_A, \mathbf{X}_B = \mathbf{x}_B)}{p(\mathbf{X}_B = \mathbf{x}_B)}$$

- This definition follows from the definition of conditional probability for events.
- Note that the numerator is the joint density and the denominator is the marginal density of $\mathbf{x}_B$.

Mixed-Type Joint Distributions

- We can also specify joint probability distributions over random variables of mixed types (e.g., discrete and continuous).
- Let $\mathbf{X} = [X_1, ..., X_{D_x}]$ be a collection of continuous random variables and $\mathbf{Y} = [Y_1, ..., Y_{D_y}]$ be a collection of discrete random variables.
- The joint distribution is typically specified as a product of a conditional and marginal distribution of one of the two following forms.

$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y})P(\mathbf{Y} = \mathbf{y})$$
$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})p(\mathbf{X} = \mathbf{x})$$

Mixed-Type Joint Distributions

- In the case shown below, the continuous random variables have a joint probability density function that is conditioned on the discrete random variables, which have a joint probability mass function.

$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y})P(\mathbf{Y} = \mathbf{y})$$

- In the case shown below, the discrete random variables have a joint probability mass function that is conditioned on the continuous random variables, which have a joint probability density function.

$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})p(\mathbf{X} = \mathbf{x})$$

Marginalization

- Suppose we have a mixed-type joint distribution where $\mathbf{X}$ is a continuous vector-valued random variable and $\mathbf{Y}$ is a discrete vector-valued random variable.
- Let $A \subseteq \{1, ..., D_x\}$ and $B = \{1, ..., D_x\} \backslash A$.
- Let $C \subseteq \{1, ..., D_y\}$ and $D = \{1, ..., D_y\} \backslash C$.
- The marginal distribution of $[\mathbf{X}_A, \mathbf{Y}_C]$ is given by:

$$
\begin{aligned}
&P(\mathbf{X}_A = x_A, \mathbf{Y}_C = \mathbf{y}_C) \\
&\quad = \int_{\mathcal{X}_B} \sum_{\mathbf{y}_D \in \mathcal{Y}_D} P(\mathbf{X}_A = \mathbf{x}_A, \mathbf{X}_B = \mathbf{x}_B, \mathbf{Y}_C = \mathbf{y}_C, \mathbf{Y}_D = \mathbf{y}_D) d\mathbf{x}_B
\end{aligned}
$$

Conditioning

- Suppose we have a mixed-type joint distribution where $\mathbf{X}$ is a continuous vector-valued random variable and $\mathbf{Y}$ is a discrete vector-valued random variable.
- Let $A \subseteq \{1, ..., D_x\}$ and $B = \{1, ..., D_x\} \backslash A$.
- Let $C \subseteq \{1, ..., D_y\}$ and $D = \{1, ..., D_y\} \backslash C$.
- The conditional distribution of $[\mathbf{X}_A, \mathbf{Y}_C]$ given $[\mathbf{X}_B, \mathbf{Y}_D]$ is:

$$
\begin{aligned}
&P(\mathbf{X}_A = x_A, \mathbf{Y}_C = \mathbf{y}_C | \mathbf{X}_B = \mathbf{x}_B, \mathbf{Y}_D = \mathbf{y}_D) \\
&= \frac{P(\mathbf{X}_A = \mathbf{x}_A, \mathbf{X}_B = \mathbf{x}_B, \mathbf{Y}_C = \mathbf{y}_C, \mathbf{Y}_D = \mathbf{y}_D)}{P(\mathbf{X}_B = \mathbf{x}_B, \mathbf{Y}_D = \mathbf{y}_D)}
\end{aligned}
$$