10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

MLE/MAP

Matt Gormley
Lecture 20
Oct 29, 2018

Q&A

PROBABILISTIC LEARNING

Probabilistic Learning

Function Approximation

Previously, we assumed that our output was generated using a **deterministic target function:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} = c^*(\mathbf{x}^{(i)})$$

Our goal was to learn a hypothesis h(**x**) that best approximates c^*(**x**)

Probabilistic Learning

Today, we assume that our output is **sampled** from a conditional **probability distribution:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} \sim p^*(\cdot|\mathbf{x}^{(i)})$$

Our goal is to learn a probability distribution p(y|**x**) that best approximates p^*(y|**x**)

Robotic Farming

	Deterministic	**Probabilistic**
Classification (binary output)	Is this a picture of a wheat kernel?	Is this plant drought resistant?
Regression (continuous output)	How many wheat kernels are in this picture?	What will the yield of this plant be?

Oracles and Sampling

Whiteboard

- Sampling from common probability distributions
  - Bernoulli
  - Categorical
  - Uniform
  - Gaussian
- Pretending to be an Oracle (Regression)
  - Case 1: Deterministic outputs
  - Case 2: Probabilistic outputs
- Probabilistic Interpretation of Linear Regression
  - Adding Gaussian noise to linear function
  - Sampling from the noise model
- Pretending to be an Oracle (Classification)
  - Case 1: Deterministic labels
  - Case 2: Probabilistic outputs (Logistic Regression)
  - Case 3: Probabilistic outputs (Gaussian Naïve Bayes)

In-Class Exercise

1. With your neighbor, **write a function** which returns **samples from a Categorical**
   - Assume access to the `rand()` function
   - Function signature should be:
     `categorical_sample(theta)`
     where theta is the array of parameters
   - Make your implementation as **efficient** as possible!
2. What is the **expected runtime** of your function?

Generative vs. Discrminative

Whiteboard

- Generative vs. Discriminative Models
  - Chain rule of probability
  - Maximum (Conditional) Likelihood Estimation for Discriminative models
  - Maximum Likelihood Estimation for Generative models

Categorical Distribution

Whiteboard

- Categorical distribution details
  - Independent and Identically Distributed (i.i.d.)
  - Example: Dice Rolls

Takeaways

- One view of what ML is trying to accomplish is **function approximation**
- The principle of **maximum likelihood estimation** provides an alternate view of learning

- **Synthetic data** can help **debug** ML algorithms
- Probability distributions can be used to **model** real data that occurs in the world
  (don’t worry we’ll make our distributions more interesting soon!)

Learning Objectives

Oracles, Sampling, Generative vs. Discriminative

You should be able to…

1. Sample from common probability distributions
2. Write a generative story for a generative or discriminative classification or regression model
3. Pretend to be a data generating oracle
4. Provide a probabilistic interpretation of linear regression
5. Use the chain rule of probability to contrast generative vs. discriminative modeling
6. Define maximum likelihood estimation (MLE) and maximum conditional likelihood estimation (MCLE)

PROBABILITY

Random Variables: Definitions

Discrete Random Variable	X	Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})
Probability mass function (pmf)	$p(x)$	Function giving the probability that discrete r.v. X takes value x. $p(x) := P(X = x)$

Random Variables: Definitions

Continuous Random Variable	X	Random variable whose values come from an interval or collection of intervals (e.g. the real numbers or the range (3, 5))
Probability density function (pdf)	$f(x)$	Function the returns a nonnegative real indicating the relative likelihood that a continuous r.v. X takes value x

- For any continuous random variable: $P(X = x) = 0$
- Non-zero probabilities are only available to intervals:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Random Variables: Definitions

Cumulative distribution function	$F(x)$	Function that returns the probability that a random variable X is less than or equal to x: $$F(x) = P(X \leq x)$$

- For **discrete** random variables:

$$F(x) = P(X \leq x) = \sum_{x' < x} P(X = x') = \sum_{x' < x} p(x')$$

- For **continuous** random variables:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x')dx'$$

Notational Shortcuts

A convenient shorthand:

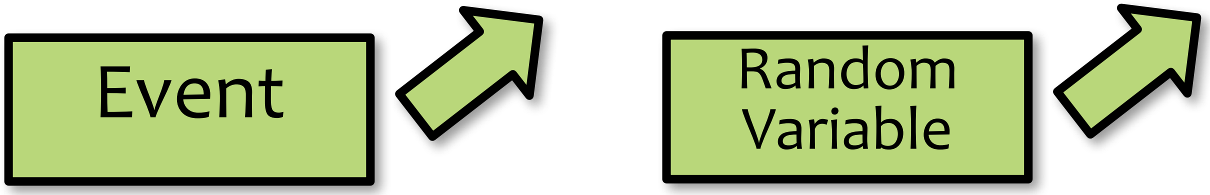
$$P(A|B) = \frac{P(A, B)}{P(B)}$$

$\Rightarrow$ For all values of a and b:

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

Notational Shortcuts

But then how do we tell $P(E)$ apart from $P(X)$?



Event

Random Variable

Instead of writing:

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

We should write:

$$P_{A|B}(A|B) = \frac{P_{A,B}(A,B)}{P_B(B)}$$

…but only probability theory textbooks go to such lengths.

COMMON PROBABILITY DISTRIBUTIONS

Common Probability Distributions

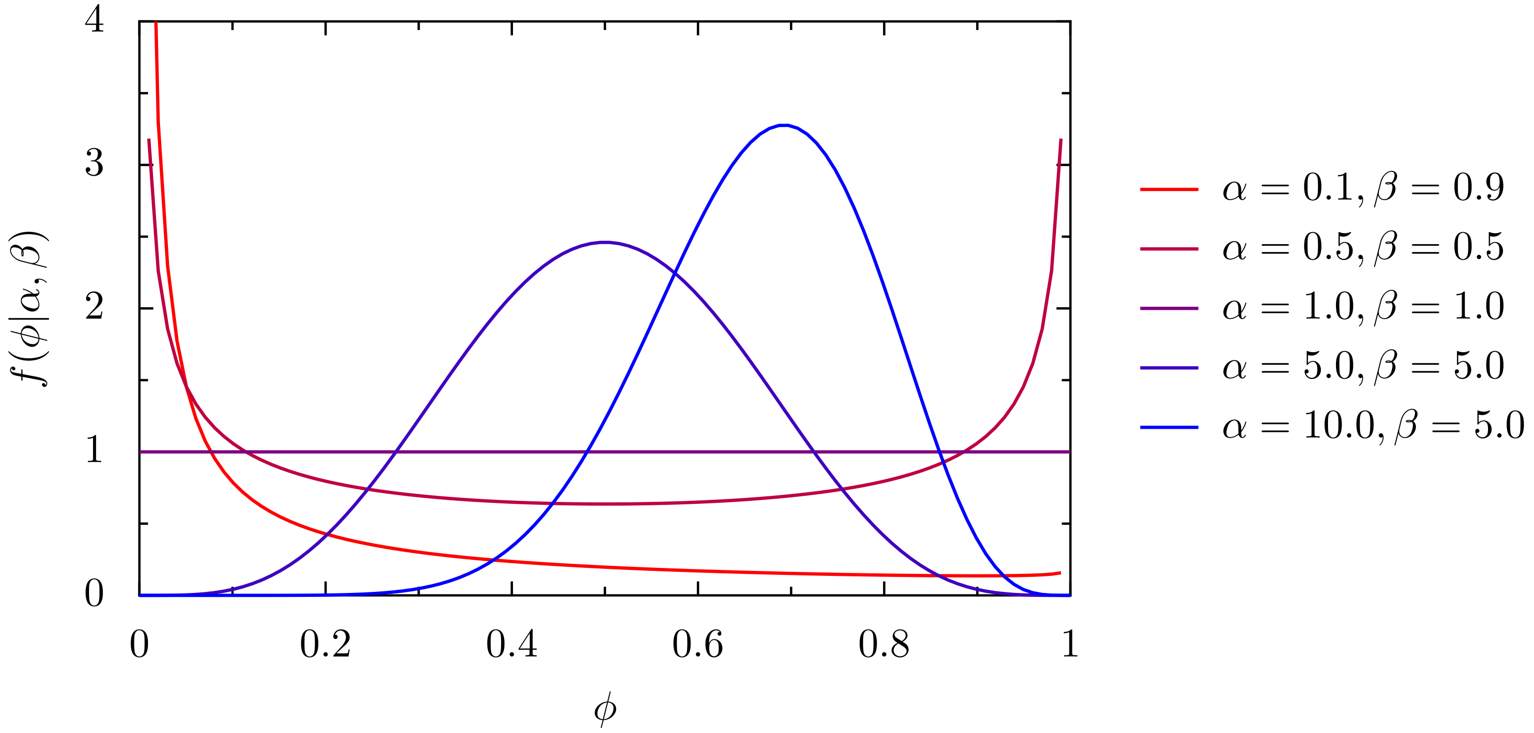
- For Discrete Random Variables:
  - Bernoulli
  - Binomial
  - Multinomial
  - Categorical
  - Poisson
- For Continuous Random Variables:
  - Exponential
  - Gamma
  - Beta
  - Dirichlet
  - Laplace
  - Gaussian (1D)
  - Multivariate Gaussian

Common Probability Distributions

Beta Distribution

probability density function:

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

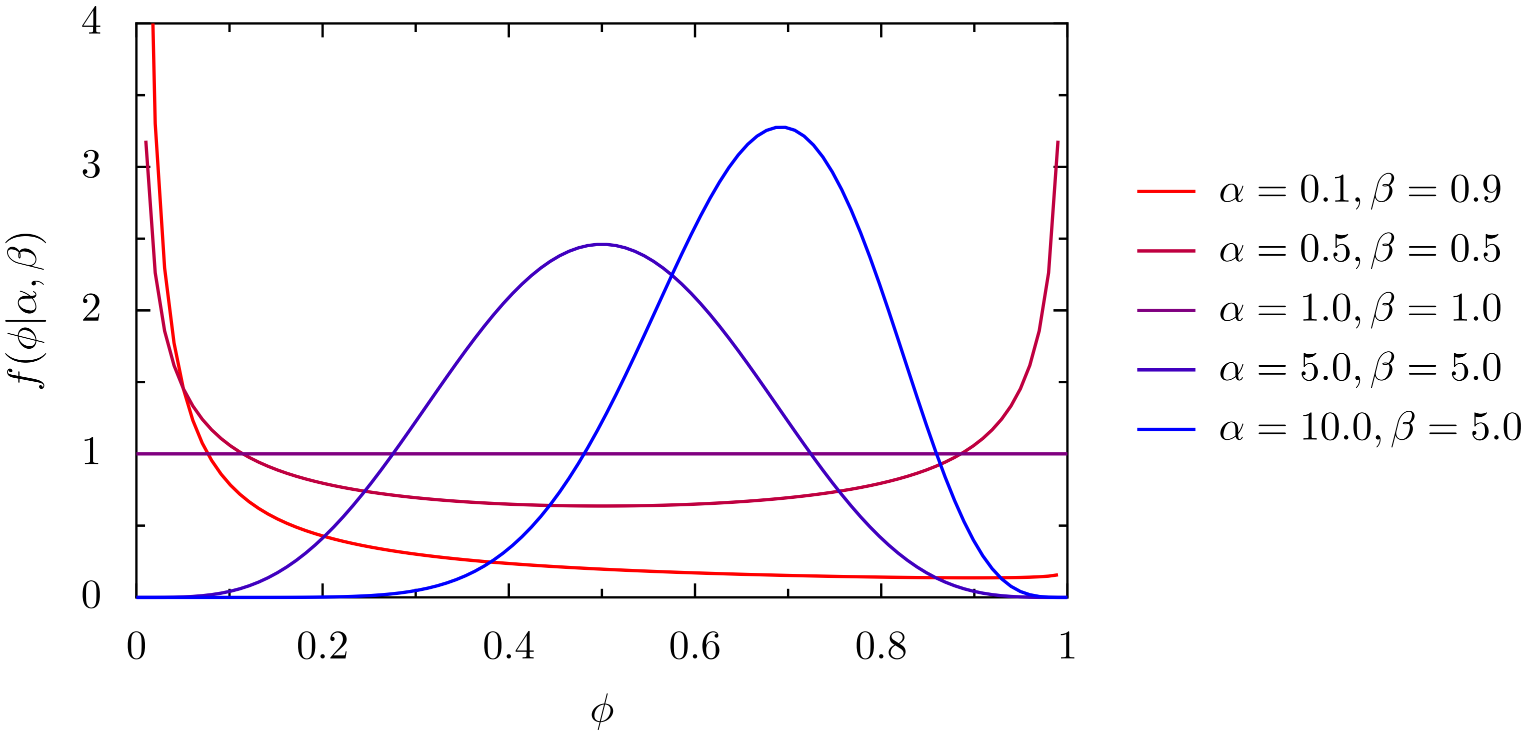


Common Probability Distributions

Dirichlet Distribution

probability density function:

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

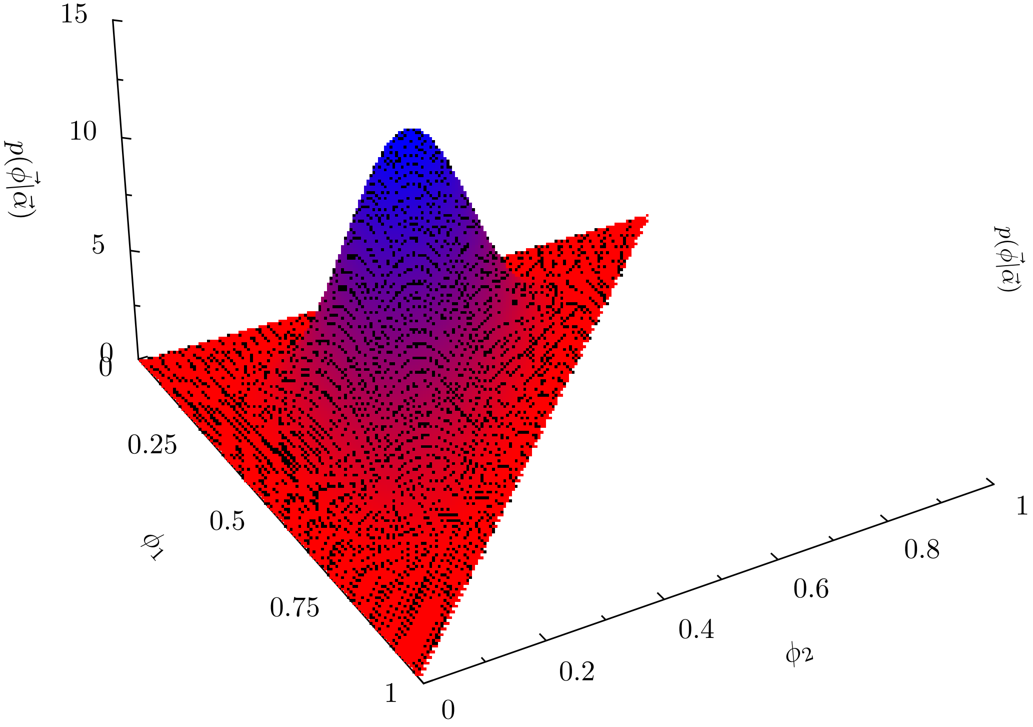


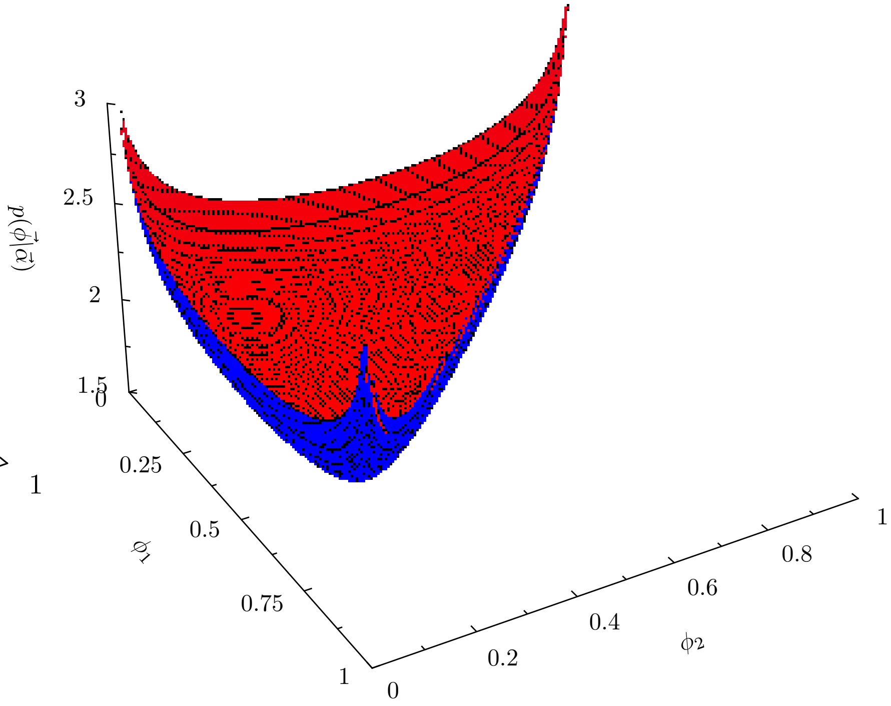
Common Probability Distributions

Dirichlet Distribution

probability density function:

$$p(\vec{\phi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \quad \text{where } B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$





EXPECTATION AND VARIANCE

Expectation and Variance

The **expected value** of X is $E[X]$. Also called the mean.

- Discrete random variables:

  Suppose X can take any value in the set $\mathcal{X}$.

  $$E[X] = \sum_{x \in \mathcal{X}} x p(x)$$

- Continuous random variables:

  $$E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

Expectation and Variance

The **variance** of *X* is *Var(X)*.

$$Var(X) = E[(X - E[X])^2]$$

$$\mu = E[X]$$

- Discrete random variables:

$$Var(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x)$$

- Continuous random variables:

$$Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Joint probability

Marginal probability

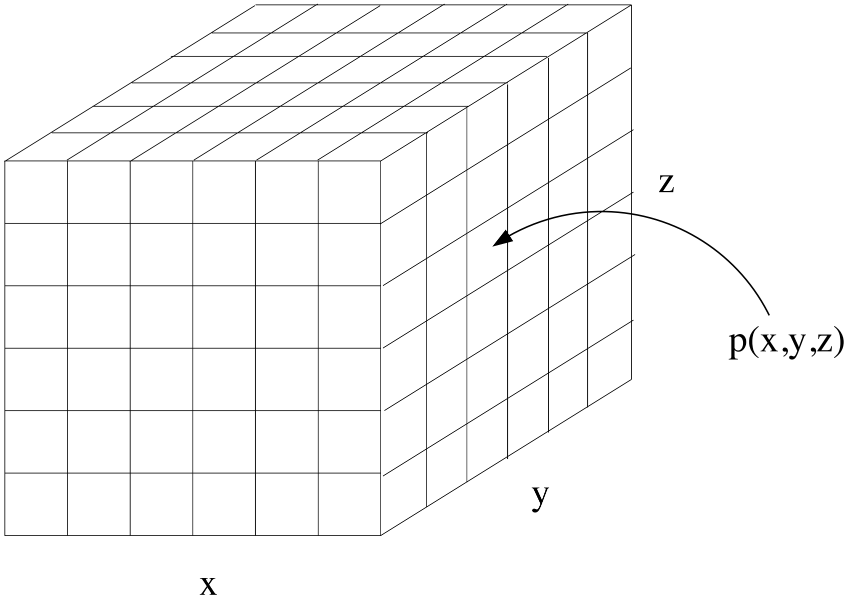
Conditional probability

MULTIPLE RANDOM VARIABLES

Joint Probability

- Key concept: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write

$$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$$



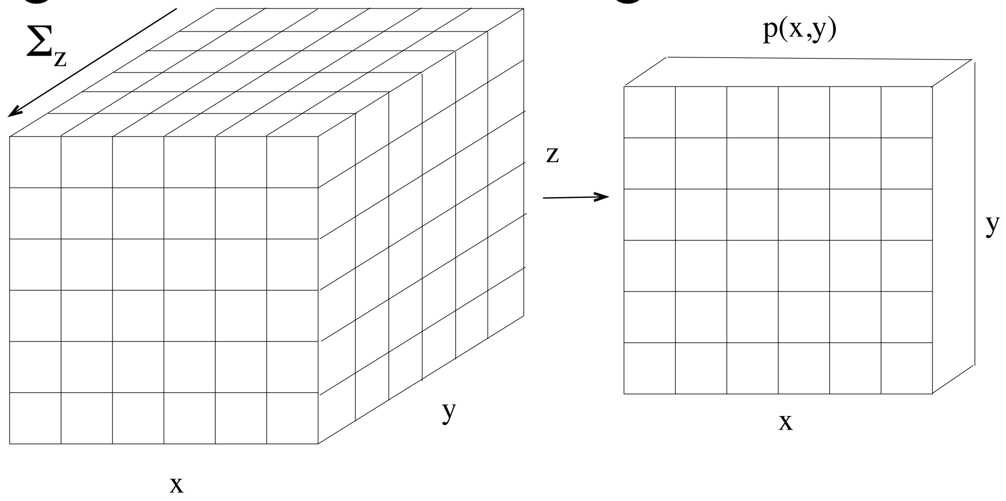
Slide from Sam Roweis (MLSS, 2005)

Marginal Probabilities

- We can "sum out" part of a joint distribution to get the *marginal distribution* of a subset of variables:

$$p(x) = \sum_y p(x, y)$$

- This is like adding slices of the table together.



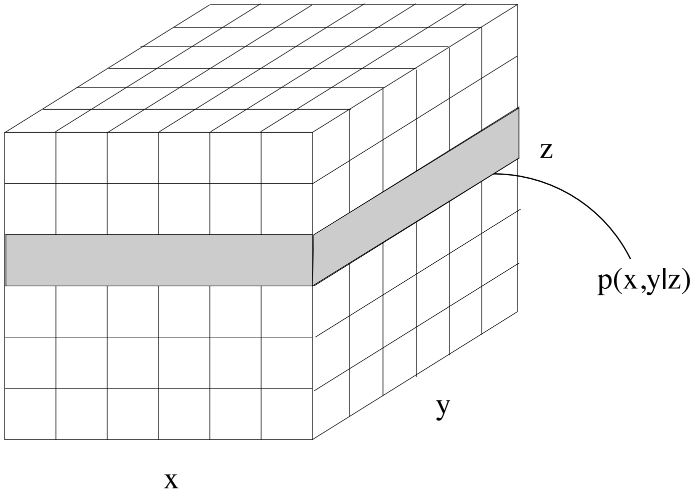
- Another equivalent definition: $p(x) = \sum_y p(x|y)p(y)$.

Slide from Sam Roweis (MLSS, 2005)

Conditional Probability

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a "slice" through the joint table.

$$\boxed{p(x|y) = p(x, y)/p(y)}$$

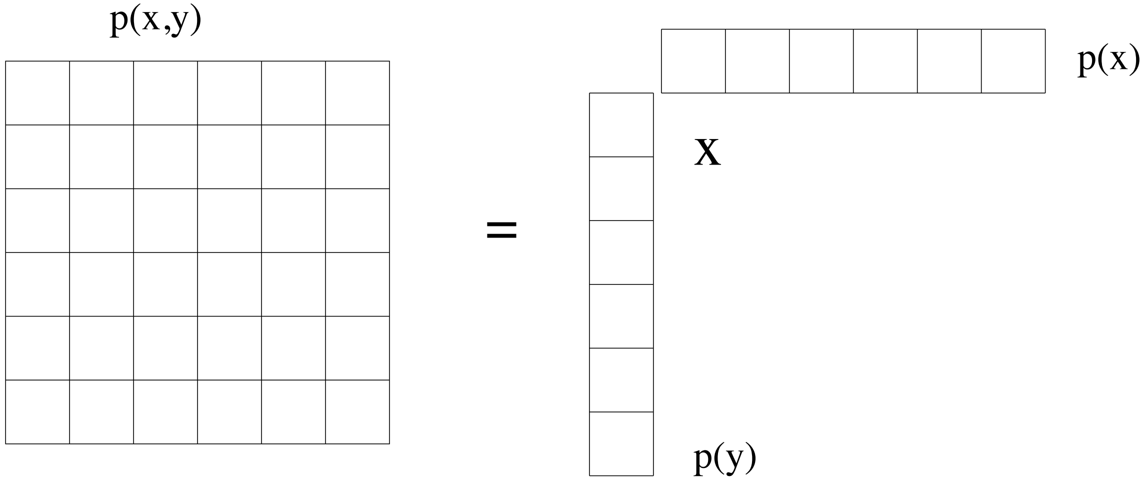


Slide from Sam Roweis (MLSS, 2005)

Independence and Conditional Independence

- Two variables are independent iff their joint factors:

$$p(x, y) = p(x)p(y)$$



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

$$p(x, y|z) = p(x|z)p(y|z) \qquad \forall z$$

Slide from Sam Roweis (MLSS, 2005)

MLE AND MAP

MLE

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

MLE

What does maximizing likelihood accomplish?

- There is only a finite amount of probability mass (i.e. sum-to-one constraint)
- MLE tries to allocate **as much** probability mass **as possible** to the things we have observed…

  …**at the expense** of the things we have **not** observed

MLE

Example: MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$
- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for λ.
- Compute second derivative and check that it is concave down at λ^{MLE}.

MLE

Example: MLE of Exponential Distribution

- First write down log-likelihood of sample.

$$
\begin{aligned}
\ell(\lambda) &= \sum_{i=1}^{N} \log f(x^{(i)}) && (1)\\
&= \sum_{i=1}^{N} \log(\lambda \exp(-\lambda x^{(i)})) && (2)\\
&= \sum_{i=1}^{N} \log(\lambda) + -\lambda x^{(i)} && (3)\\
&= N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} && (4)
\end{aligned}
$$

MLE

Example: MLE of Exponential Distribution

- Compute first derivative, set to zero, solve for λ.

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda} N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \tag{1}$$

$$= \frac{N}{\lambda} - \sum_{i=1}^{N} x^{(i)} = 0 \tag{2}$$

$$\boxed{\Rightarrow \lambda^{\text{MLE}} = \frac{N}{\sum_{i=1}^{N} x^{(i)}}} \tag{3}$$

MLE

Example: MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$
- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for λ.
- Compute second derivative and check that it is concave down at λ^{MLE}.

MLE

In-Class Exercise

Show that the MLE of parameter ϕ for N samples drawn from Bernoulli(ϕ) is:

$$\phi_{MLE} = \frac{\text{Number of } x_i = 1}{N}$$

Steps to answer:

1. Write log-likelihood of sample
2. Compute derivative w.r.t. ϕ
3. Set derivative to zero and solve for ϕ

Learning from Data (Frequentist)

Whiteboard

- – Optimization for MLE
- – Examples: 1D and 2D optimization
- – Example: MLE of Bernoulli
- – Example: MLE of Categorical
- – Aside: Method of Langrange Multipliers

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

Principle of Maximum *a posteriori* (MAP) Estimation:
Choose the parameters that maximize the posterior of the parameters given the data.

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\boldsymbol{\theta}|\mathbf{x}^{(i)})$$

Maximum *a posteriori* (MAP) estimate

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

Principle of Maximum *a posteriori* (MAP) Estimation:
Choose the parameters that maximize the posterior of the parameters given the data.

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})\overbrace{p(\boldsymbol{\theta})}^{\text{Prior}}$$

Maximum *a posteriori* (MAP) estimate

Learning from Data (Bayesian)

Whiteboard

- *maximum a posteriori* (MAP) estimation
- Optimization for MAP
- Example: MAP of Bernoulli—Beta

Takeaways

- One view of what ML is trying to accomplish is **function approximation**
- The principle of **maximum likelihood estimation** provides an alternate view of learning

- **Synthetic data** can help **debug** ML algorithms
- Probability distributions can be used to **model** real data that occurs in the world
  (don't worry we'll make our distributions more interesting soon!)

Learning Objectives

MLE / MAP

You should be able to…

1. Recall probability basics, including but not limited to: discrete and continuous random variables, probability mass functions, probability density functions, events vs. random variables, expectation and variance, joint probability distributions, marginal probabilities, conditional probabilities, independence, conditional independence
2. Describe common probability distributions such as the Beta, Dirichlet, Multinomial, Categorical, Gaussian, Exponential, etc.
3. State the principle of maximum likelihood estimation and explain what it tries to accomplish
4. State the principle of maximum a posteriori estimation and explain why we use it
5. Derive the MLE or MAP parameters of a simple model in closed form