THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

CS 4364/6364

Machine Learning

Fall Semester 8/31/2023
Lecture 3.
Probability and Information Theory

John Sipple
jsipple@gwu.edu

THE WALL STREET JOURNAL.

John Sipple ▾

MARKETS NEWSLETTER

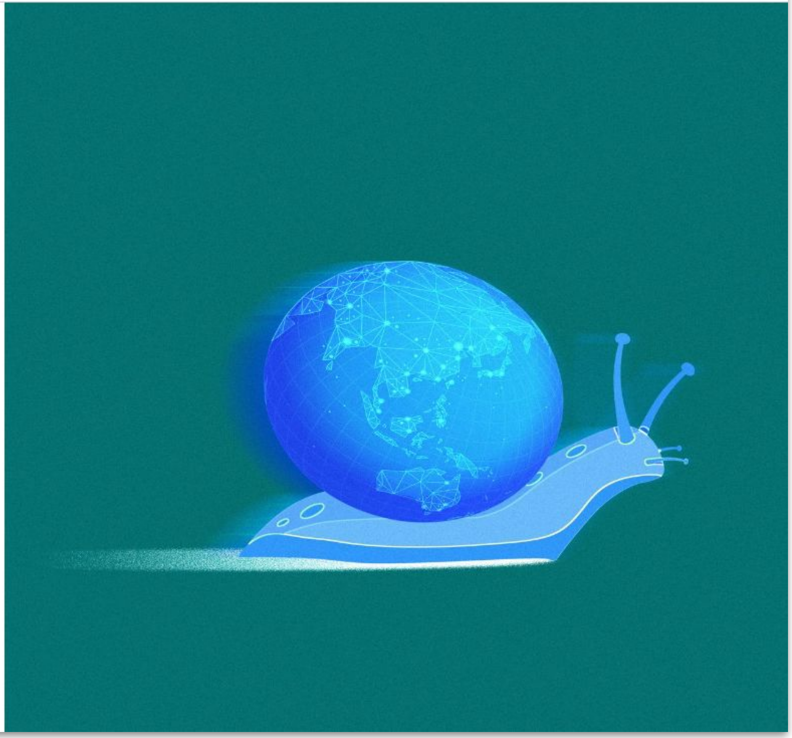
English Edition ▾ | Print Edition | Video | Audio | Latest Headlines | More ▾

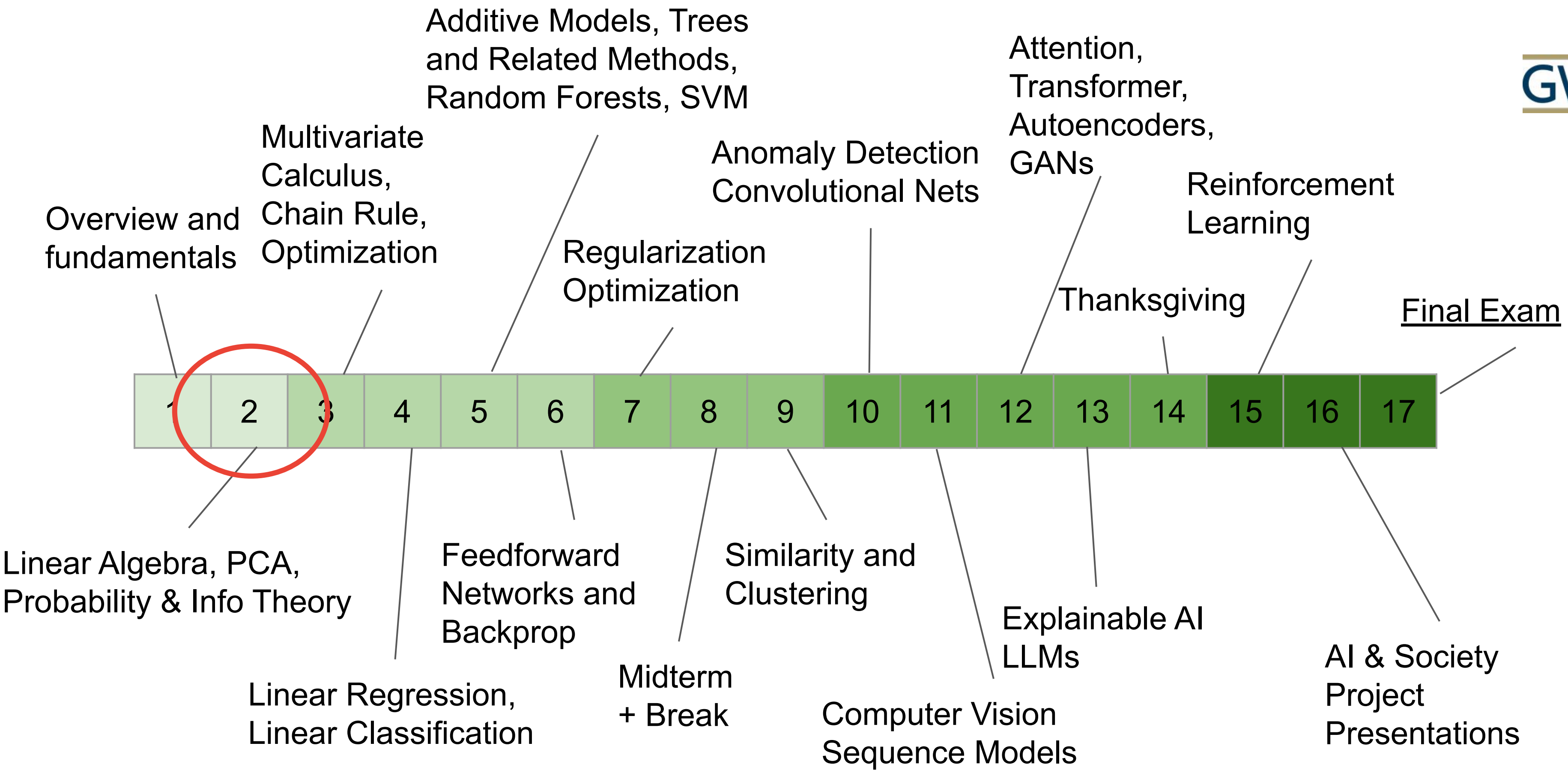
World Business U.S. Politics Economy Tech Finance Opinion Arts & Culture Lifestyle Real Estate Personal Finance Health Science Style Sports

TECHNOLOGY | ARTIFICIAL INTELLIGENCE

AI Startup Buzz Is Facing a Reality Check

Venture investors are realizing that generative artificial intelligence might not be enough to stem yearslong startup downturn




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Overview and fundamentals
Linear Algebra, PCA, Probability & Info Theory
Multivariate Calculus, Chain Rule, Optimization
Linear Regression, Linear Classification
Additive Models, Trees and Related Methods, Random Forests, SVM
Feedforward Networks and Backprop
Regularization Optimization
Midterm + Break
Similarity and Clustering
Anomaly Detection Convolutional Nets
Computer Vision Sequence Models
Attention, Transformer, Autoencoders, GANs
Explainable AI LLMs
Thanksgiving
Reinforcement Learning
AI & Society Project Presentations
Final Exam

Probability

<u>Language of uncertainty</u>

x - Random Variable (RV), that can take on a value in its domain

$\mathrm{x} = x$, An Event where RV x takes on a value x

Degree of belief that x will occur given that we know y has occurred:

$P(\mathrm{x}=x|\mathrm{y}=y) \in [0, 1]$

Probability Mass Function

The domain of P must be the set of all possible states (events) of $\mathbf{x}$.

$$\forall x \in \mathbf{x}, 0 \leq P(x) \leq 1$$

From $P(\text{Impossible Event}) = 0$ to $P(\text{Completely Certain Event}) = 1$

The PMF is normalized, ensuring that it bounded between 0 and 1: $\sum_{x \in \mathbf{x}} P(x) = 1$

Example: Uniform distribution with k possible events: $P(\mathbf{x} = x_i) = \frac{1}{k}$

Probability Density Function

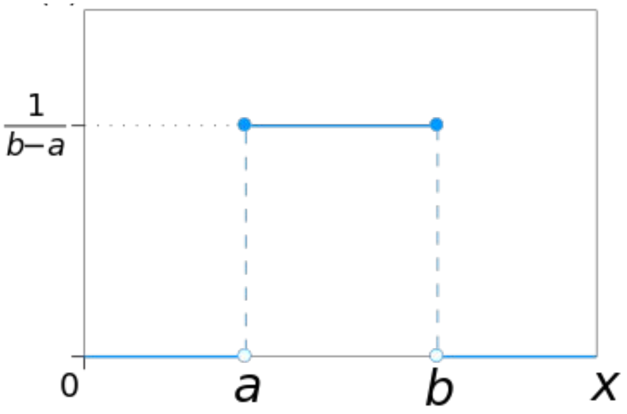
The domain if p must be the set of all possible states of x.

$\forall x \in \mathbf{x}, p(x) \geq 0$ Note that we do not require $p(x) \leq 1$

$\int p(x)dx = 1$

Example: uniform distribution:

$$u(x; a, b) = \frac{1}{b-a}$$



https://en.wikipedia.org/wiki/Continuous_uniform_distribution

Marginal Probability

$$\forall x \in \mathbf{x}, P(\mathbf{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

$$p(x) = \int p(x, y) dy$$

Marginal Probability (example)

		s		
		Male	Female	Total
a	Football	120	75	195
	Rugby	100	25	125
	Other	50	130	180
		270	230	500

Adapted from: Marginal, Joint and Conditional Probabilities explained By Data Scientist
https://towardsdatascience.com/marginal-joint-and-conditional-probabilities-explained-by-data-scientist-4225b28907a4

Marginal Probability (example)

s

a

	Male	Female	Total
Football	0.24	0.15	0.39
Rugby	0.2	0.05	0.25
Other	0.1	0.26	0.36
	0.54	0.46	1

Marginal Probability (example)

		s	
a	Male	Female	Total
Football	0.24	0.15	0.39
Rugby	0.2	0.05	0.25
Other	0.1	0.26	0.36
	0.54	0.46	1

Joint Probability $P(\mathbf{s},\mathbf{a})$

$$P(\mathbf{s} = Female, \mathbf{a} = Rugby) = 0.05$$

Marginal Probability (example)

		s		
		Male	Female	Total
a	Football	0.24	0.15	0.39
	Rugby	0.2	0.05	0.25
	Other	0.1	0.26	0.36
		0.54	0.46	1

Marginal Probabilities
$P(\mathbf{s}), P(\mathbf{a})$

$$P(s = Female)$$
$$= P(s = Female, a = Football) + P(s = Female, a = Rugby) + P(s = Female, a = Other)$$
$$= 0.46$$

Joint and Conditional Probability

$$\begin{aligned} &P(\mathrm{x}=x,\mathrm{y}=y) \\ &= P(\mathrm{y}=y|\mathrm{x}=x)P(\mathrm{x}=x) \\ &= P(\mathrm{x}=x|\mathrm{y}=y)P(\mathrm{y}=y) \end{aligned}$$

$$P(\mathrm{y}=y|\mathrm{x}=x)=\frac{P(\mathrm{y}=y,\mathrm{x}=x)}{P(\mathrm{x}=x)}$$

Conditional Probability (example)

$\mathbf{s}$

$\mathbf{a}$

	Male	Female	Total
Football	0.24	0.15	0.39
Rugby	0.2	0.05	0.25
Other	0.1	0.26	0.36
	0.54	0.46	1

Conditional Probabilities

$P(\mathbf{s}|\mathbf{a}), P(\mathbf{a}|\mathbf{s})$

$$P(\mathbf{a} = Football|\mathbf{s} = Male) = \frac{P(\mathbf{a} = Football, \mathbf{s} = Male)}{P(\mathbf{s} = Male)} = \frac{0.24}{0.54} = 0.44$$

Chain Rule of Probability

Joint Probability for variables $x^{(1)}, x^{(2)}, x^{(3)}$

$$P(\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \mathrm{x}^{(3)}) = P(\mathrm{x}^{(1)}) P(\mathrm{x}^{(2)} | \mathrm{x}^{(1)}) P(\mathrm{x}^{(3)} | \mathrm{x}^{(2)}, \mathrm{x}^{(1)})$$

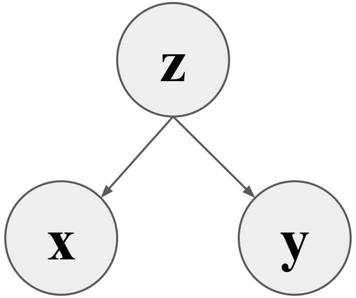
Joint Probability for n variables $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \prod_{i=2}^{n} P(\mathrm{x}^{(i)} | \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y},$$
$$p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$$

Conditional Independence



$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},$$
$$p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)\, p(\mathrm{y} = y \mid \mathrm{z} = z)$$

Expectation

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x)f(x)$$

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx$$

Linearity of expectations:

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}_{\mathrm{x}}[f(x)] + \beta\mathbb{E}_{\mathrm{x}}[g(x)]$$

Variance and Covariance

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

$$\mathrm{Cov}(f(x), g(x)) = \mathbb{E}[f(x) - \mathbb{E}(f(x)(g(x) - \mathbb{E}(g(x))]$$

Covariance Matrix

$$\Sigma = \mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j)$$

Bernoulli Distribution

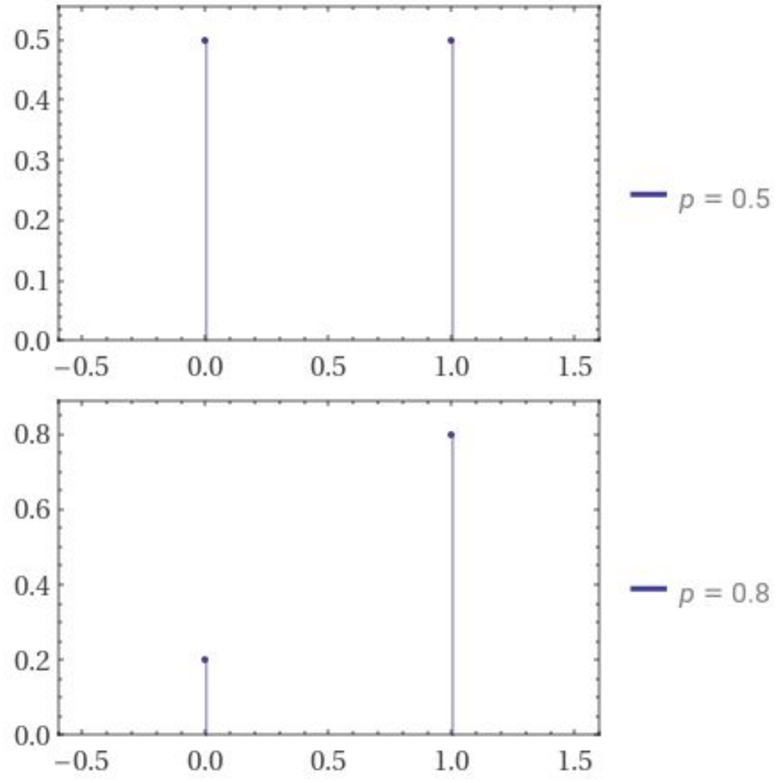
$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^x (1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$



Gaussian Distribution

Parameterized by variance

$$\mathcal{N}(x;\mu,\sigma^2)=\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Parameterized by precision:

$$\mathcal{N}(x;\mu,\beta^{-1})=\sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{\beta}{2}(x-\mu)^2\right)$$

Gaussian Distribution

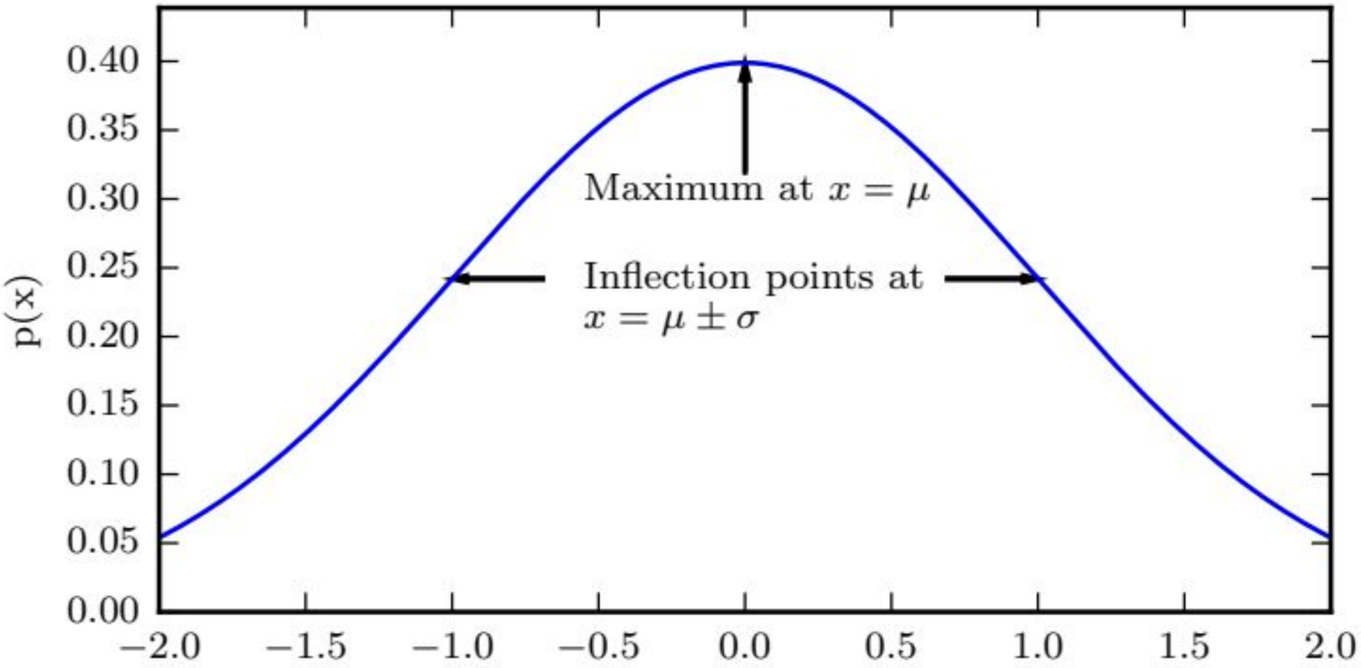


Figure 3.1

Multivariate Gaussian

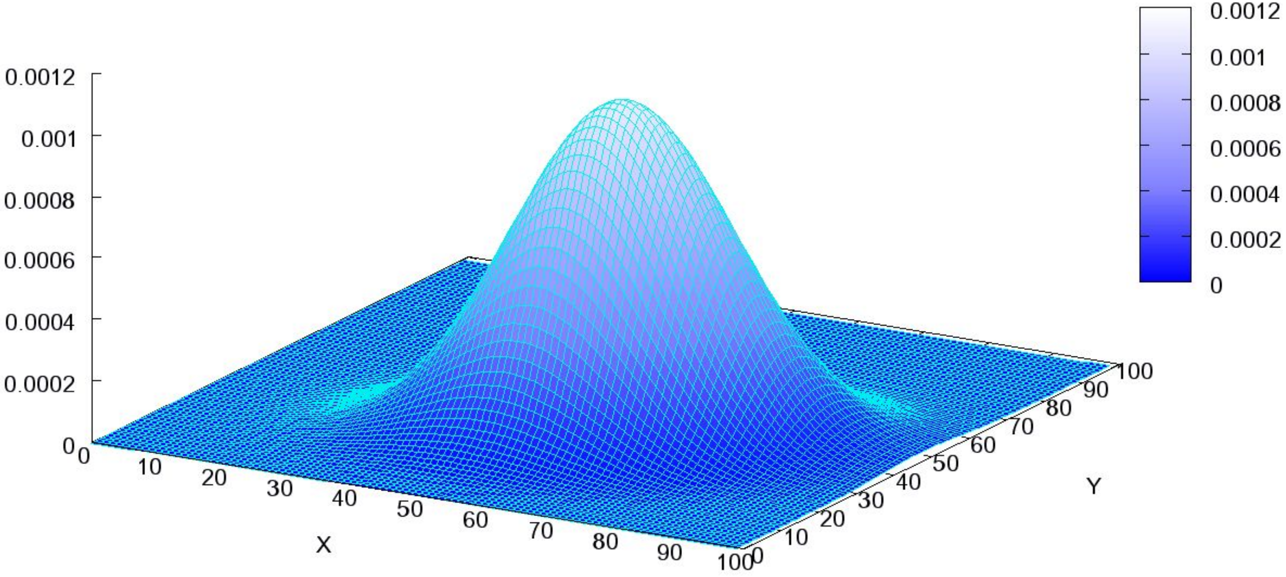
Parameterized by covariance matrix Σ:

$$\mathcal{N}(\mathbf{x};\mu,\Sigma) = \sqrt{\frac{1}{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

Parameterized by precision matrix β:

$$\mathcal{N}(\mathbf{x};\mu,\beta^{-1}) = \sqrt{\frac{\det(\beta)}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathbf{T}} \beta(\mathbf{x}-\mu)\right)$$

Multivariate Gaussian

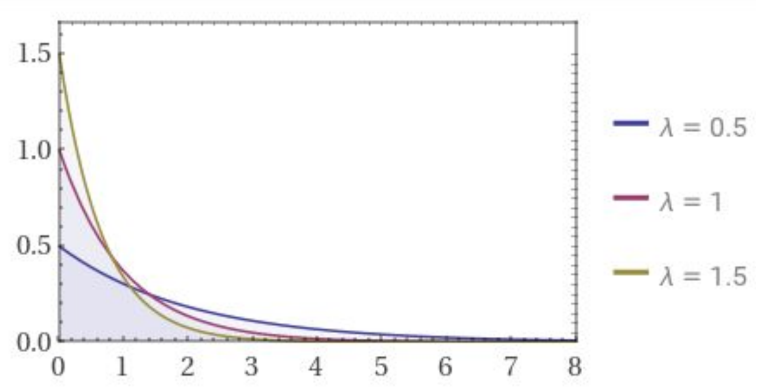


Exponential Distribution

Long-tailed distribution

Often represents occurrence counts of some process

λ controls the shape of the distribution



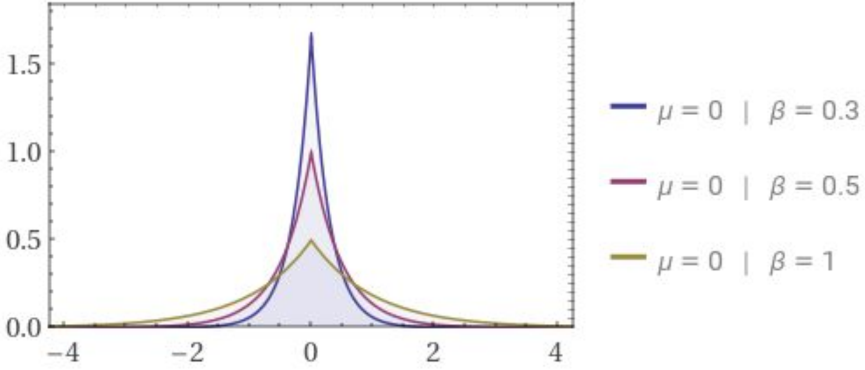
$$p(x;\lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x)$$

Laplace Distribution

Difference of two independent exponential variables

Peak at the mean μ

γ controls the shape of the distribution



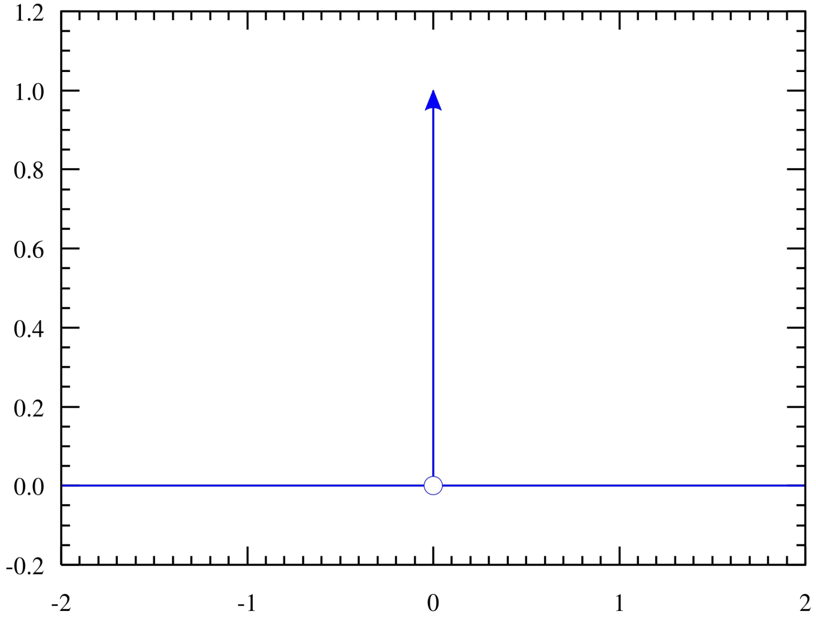
$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$$

Dirac Delta Function

"Unit Impulse Function"

1 at parameter μ, 0 everywhere else

$$p(x) = \delta(x - \mu)$$



Empirical Distribution

Of m observed data points,
each point $x^{(i)}$ gets weight $1/m$

$$\hat{p}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\mathbf{x} - \mathbf{x}^{(i)})$$

Mixture Distributions

aka **Multimodal distribution**

$$P(\mathrm{x}) = \sum_i P(c = i)P(\mathrm{x}|\mathrm{c})$$

Gaussian mixture with three components

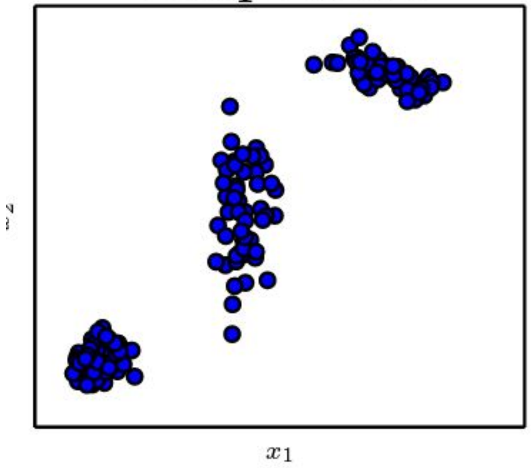


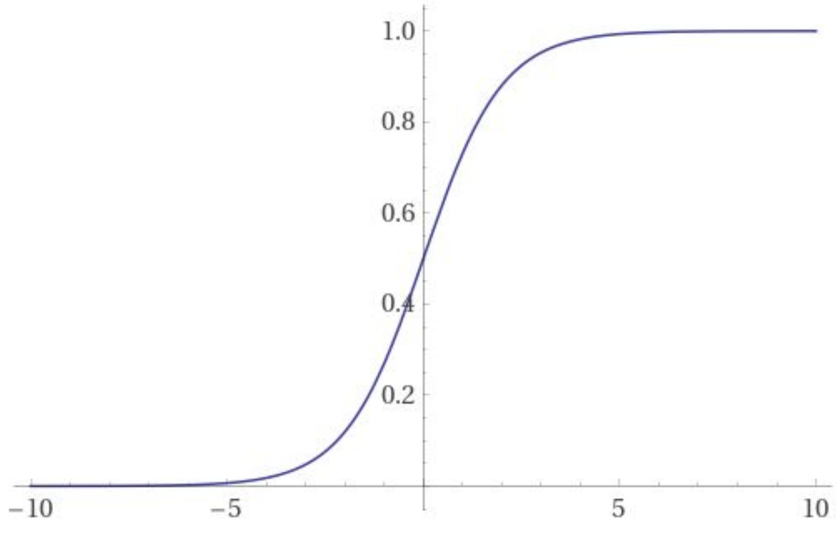
Figure 3.2

Logistic Sigmoid

Forces x (aka *logit*) into range of [0,1]:

Used in logistic regression, DNN binary classifiers to approximate probability

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$



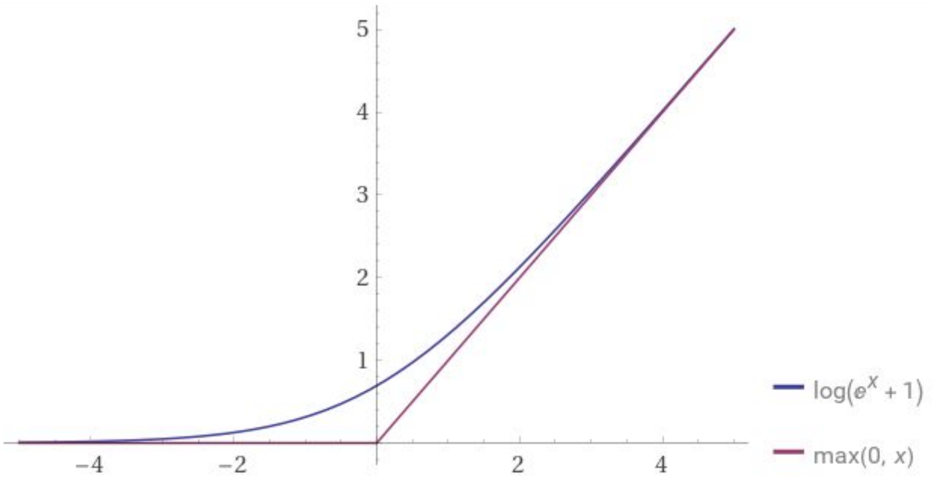
Soft Plus Function

Smoothed version of

$$x^+ = \max(0, x)$$

Can be used to parameterize σ in a Normal distribution

$$\zeta(x) = \log(1 + \exp(x))$$



Bayes Rule

$$P(\mathrm{x}|\mathrm{y}) = \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{P(\mathrm{y}|\mathrm{x})P(\mathrm{x})}{P(\mathrm{y})}$$

Change of Variables

Given a continuous, invertible, continuously differentiable function

$$\boldsymbol{y} = g(\boldsymbol{x})$$

Doesn't guarantee the area under the probability curve is 1!

$$p_y(\boldsymbol{y}) \neq p_x(g^{-1}(\boldsymbol{y}))$$

Need to rescale:

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x})) \left| \det \left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}} \right) \right|$$

Information Theory

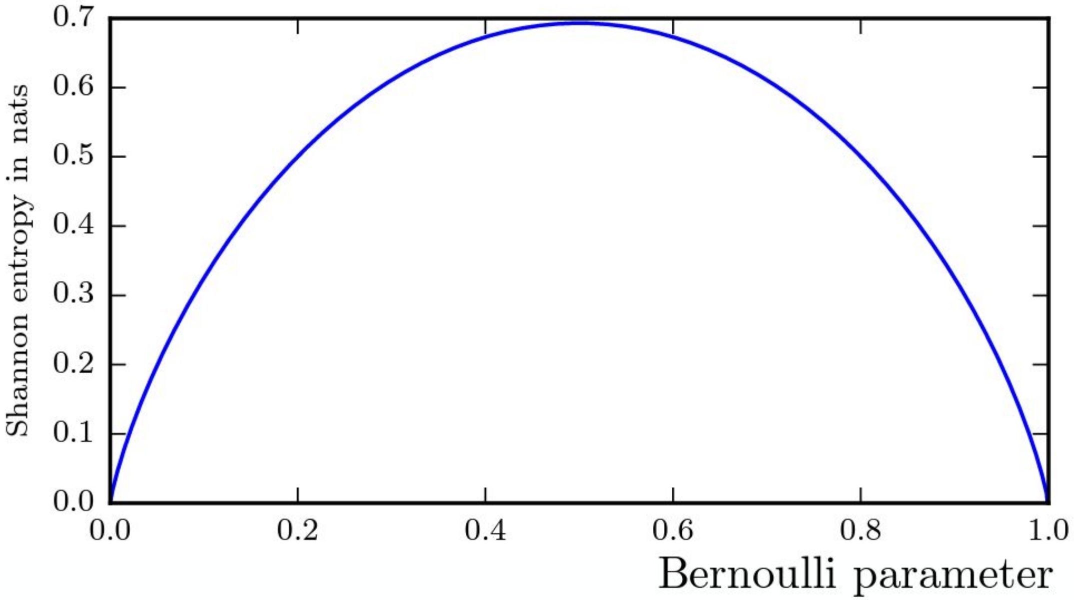
Basic Idea: Learning that an unlikely event occurred is more informative than learning that a likely event occurred.

Information of a single event: $I(x) = -\log P(\mathrm{x} = x)$

Entropy of a distribution:

$$H(\mathrm{x}) = \mathbb{E}_{x \sim P}[I(x)] = -\mathbb{E}_{x \sim P}[\log P(x)]$$
$$H(\mathrm{x}) = -\sum_{x} P(\mathrm{x} = x) \log P(\mathrm{x} = x)$$

Entropy of a Coin Flip (Bernoulli RV)



$$H(\mathrm{x}) = -(1-p)\log(p-1) - p\log(p)$$

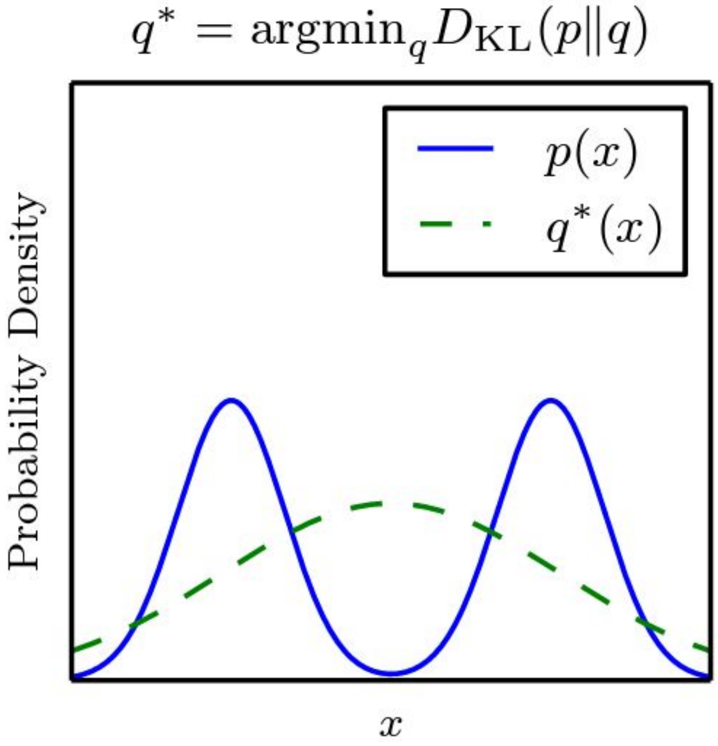
Kullback-Leibler (KL) Divergence

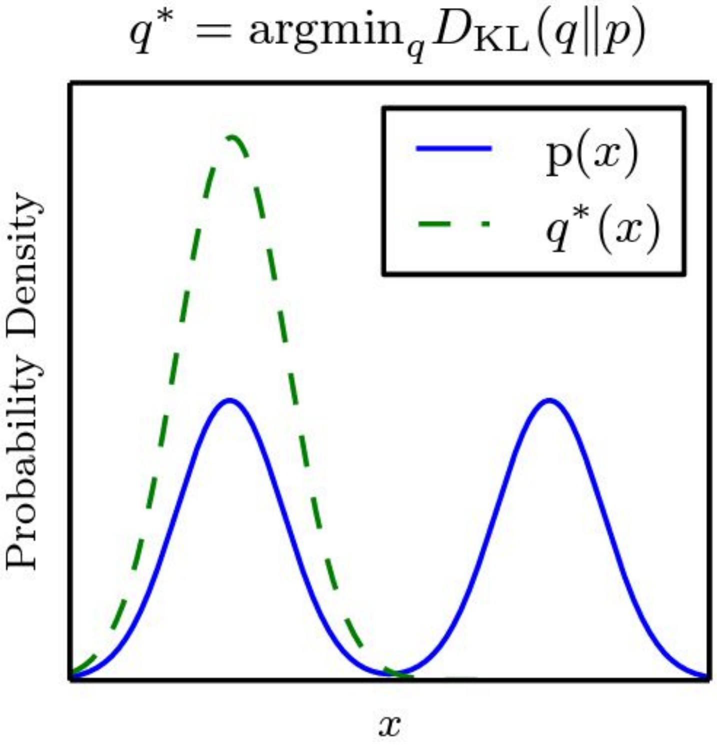
Measures the difference between two distributions from the "perspective" of one distribution

$$D_{KL}(P||Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - logQ(x)\right]$$

$$D_{KL}(P||Q) = \sum_{x} P(\mathrm{x}=x)\log \frac{P(\mathrm{x}=x)}{Q(\mathrm{x}=x)}$$

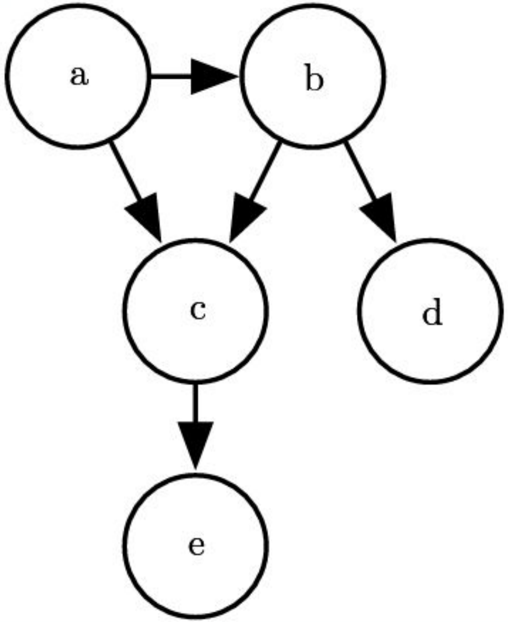
KL Divergence is Asymmetric





Directed Structured Model

Figure 3.7



$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}).$$

Undirected Structured Model

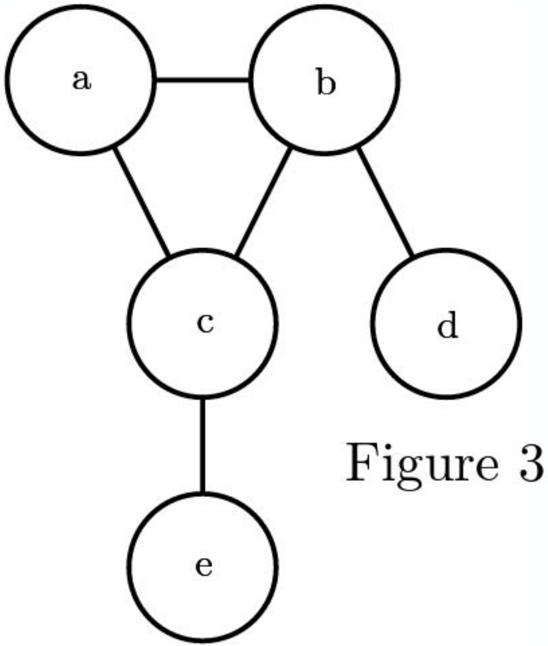


Figure 3.8

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = \frac{1}{Z}\phi^{(1)}(\mathrm{a}, \mathrm{b}, \mathrm{c})\phi^{(2)}(\mathrm{b}, \mathrm{d})\phi^{(3)}(\mathrm{c}, \mathrm{e}).$$