An Introduction to Probability Theory

Mathematics for Machine Learning

Lecturer: Matthew Wicker

Logistics

Lecture: Lectures will be shortened from the full two hours to an hour to an hour and a half with tutorial time following

Coursework: Coursework has been posted and should be accessible through LabTS

Recordings: Lectures 4 and 5 that were without audio are being re-recorded today and tomorrow and will be posted to panopto

Material Covered

Models: Linear models, basis expansion, logistic regression, neural networks

Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity, Maximum likelihood, maximum a posteriori,

Settings: Regression, Classification, Density Estimation

This lecture: Basis in probability theory, manipulating probability distributions

A Primer on Measure Theory

Definition 2.1. σ**-Algebra** A σ-Algebra on a non-empty set Ω is a collection $A \subseteq \mathcal{P}(\Omega)$ such that the collection is closed under compliments and countable unions. Formally:

1. (Closed under compliment): If a set $B \in A$, then that implies $\bar{B} \in A$, that its compliment is in A
2. (Closed under countable union): If a series of collections $B_1, B_2, \ldots, B_n \in A$, then that implies $\bigcup_{i=1}^{n} B_i \in A$

Sigma Algebra

Definition 2.1. **σ-Algebra** A σ-Algebra on a non-empty set Ω is a collection $A \subseteq \mathcal{P}(\Omega)$ such that the collection is closed under compliments and countable unions. Formally:

1. (Closed under compliment): If a set $B \in A$, then that implies $\bar{B} \in A$, that its compliment is in A
2. (Closed under countable union): If a series of collections $B_1, B_2, \ldots, B_n \in A$, then that implies $\bigcup_{i=1}^{n} B_i \in A$

$$\Omega \in A(E \in A, E^c \in A, E \cup E^c \in A, (E \cup E^c = \Omega))$$

Example: Sigma Algebra

$$A = \{\emptyset, \Omega\}$$

$$A = \{\emptyset, E, E^c, \Omega\}$$

Definition 2.1. **σ-Algebra** A σ-Algebra on a non-empty set Ω is a collection $A \subseteq \mathcal{P}(\Omega)$ such that the collection is closed under compliments and countable unions. Formally:

1. (Closed under compliment): If a set $B \in A$, then that implies $\bar{B} \in A$, that its compliment is in A
2. (Closed under countable union): If a series of collections $B_1, B_2, \ldots, B_n \in A$, then that implies $\bigcup_{i=1}^{n} B_i \in A$

Example: Sigma Algebra

$$A = \{\emptyset, \Omega\}$$

$$A = \{\emptyset, E, E^c, \Omega\}$$

$$\text{if } \Omega = \mathbb{R}, B = \sigma(\tau) \text{ where } \tau = \{(a, b)\} \forall a < b \in \mathbb{R}$$

The Borel measure

Definition 2.1. **σ-Algebra** A σ-Algebra on a non-empty set Ω is a collection $A \subseteq \mathcal{P}(\Omega)$ such that the collection is closed under compliments and countable unions. Formally:

1. (Closed under compliment): If a set $B \in A$, then that implies $\bar{B} \in A$, that its compliment is in A
2. (Closed under countable union): If a series of collections $B_1, B_2, \ldots, B_n \in A$, then that implies $\bigcup_{i=1}^{n} B_i \in A$

What is a probability measure?

Definition 2.2. Probability Measure A probability measure, P over a σ-Algebra (Ω, A) is a function $P : A \mapsto [0, \infty]$ such that:

1. $P(\emptyset) = 0$ and $P(\Omega) = 1$
2. $P(\bigcup_{i=1}^{n} B_i) = \sum_{i=1}^{n} P(B_i)$ as long as each B_i is pairwise disjoint

Example of Prob. Measure (Uniform)

$$\Omega = \{1, 2, \ldots, n\},\ A = \mathcal{P}(\Omega)$$

$$P(\{k\}) = \frac{1}{n}$$

Definition 2.2. Probability Measure A probability measure, P over a σ-Algebra (Ω, A) is a function $P : A \mapsto [0, \infty]$ such that:

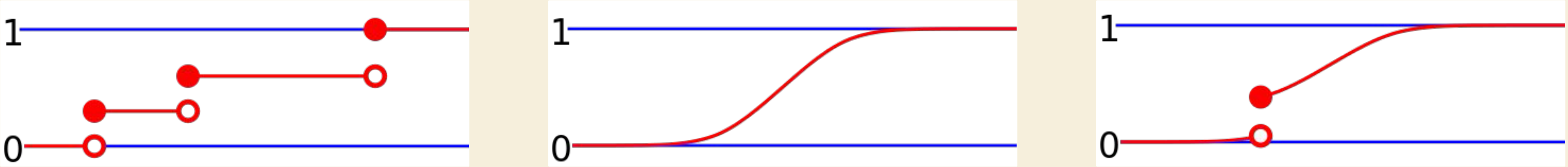
1. $P(\emptyset) = 0$ and $P(\Omega) = 1$
2. $P(\bigcup_{i=1}^{n} B_i) = \sum_{i=1}^{n} P(B_i)$ as long as each B_i is pairwise disjoint

Continuous Probability

Definition 2.3. Cumulative Distribution Function A cumulative distribution function (c.d.f. or cdf) or a probability measure is a function $F : \mathbb{R} \mapsto \mathbb{R}$ such that:

1. $x \leq y \implies F(x) < F(y)$
2. $\lim_{x \downarrow y} F(x) = F(y)$ (i.e., right continuous)
3. $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$

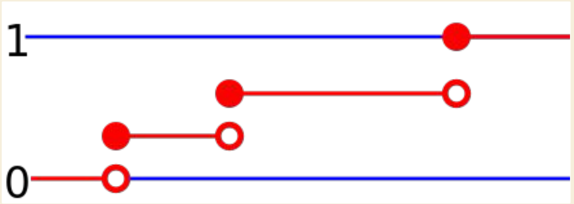
Cumulative Distribution Function

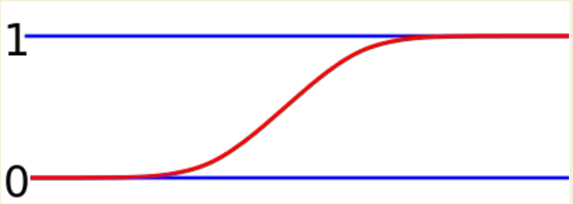


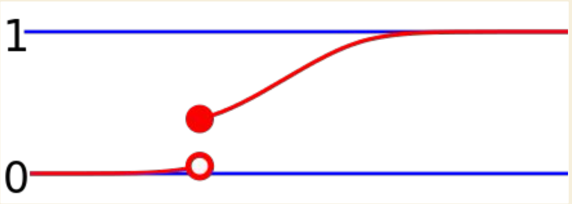
Definition 2.3. Cumulative Distribution Function A cumulative distribution function (c.d.f. or cdf) or a probability measure is a function $F : \mathbb{R} \mapsto \mathbb{R}$ such that:

1. $x < y \implies F(x) < F(y)$
2. $\lim_{x \downarrow y} F(x) = F(y)$ (i.e., right continuous)
3. $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$

Cumulative Distribution Function



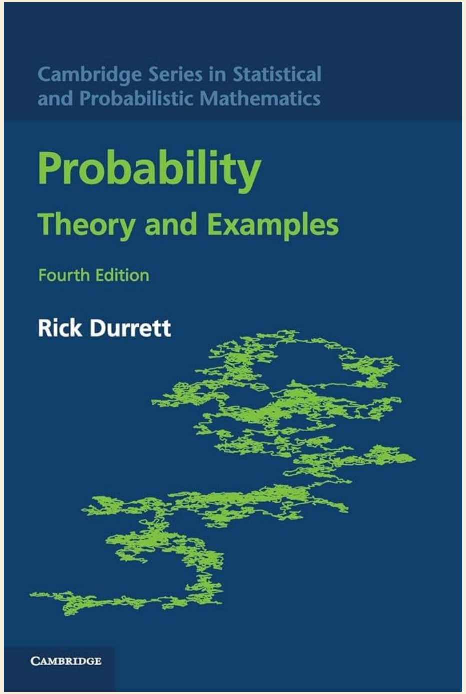




For each cumulative distribution function, there is a unique probability measure.

For each probability measure there is a unique cumulative distribution function

This is as far as we go with measure theory



https://services.math.duke.edu/·rtd/PTE/PTE5_011119.pdf

Statistics terminology

Ω - Sample space

$\sigma(\Omega)$ - Event space

P - Probability Measure

Random Variable

$$X \begin{cases} \Omega \text{ - Sample space} \\ \sigma(\Omega) \text{ - Event space} \\ P \text{ - Probability Measure} \end{cases}$$

Random Variable

Often these will be implicitly defined as we will be dealing with random variables over the reals

$$X \begin{cases} \boxed{\begin{array}{c} \Omega \text{ - Sample space} \\ \sigma(\Omega) \text{ - Event space} \end{array}} \\ P \text{ - Probability Measure} \end{cases}$$

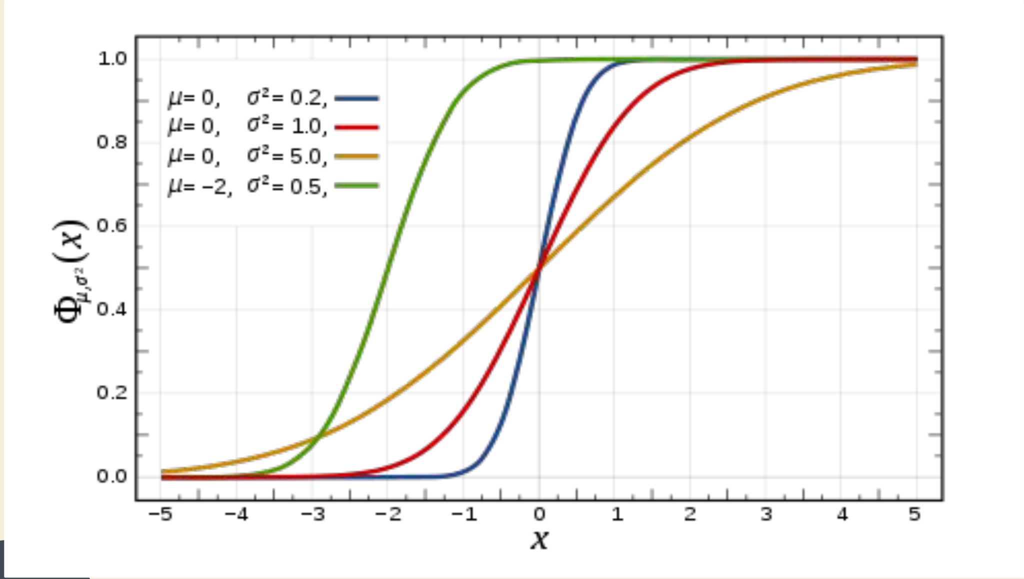
Random Variable

Often these will be implicitly defined as we will be dealing with random variables over the reals

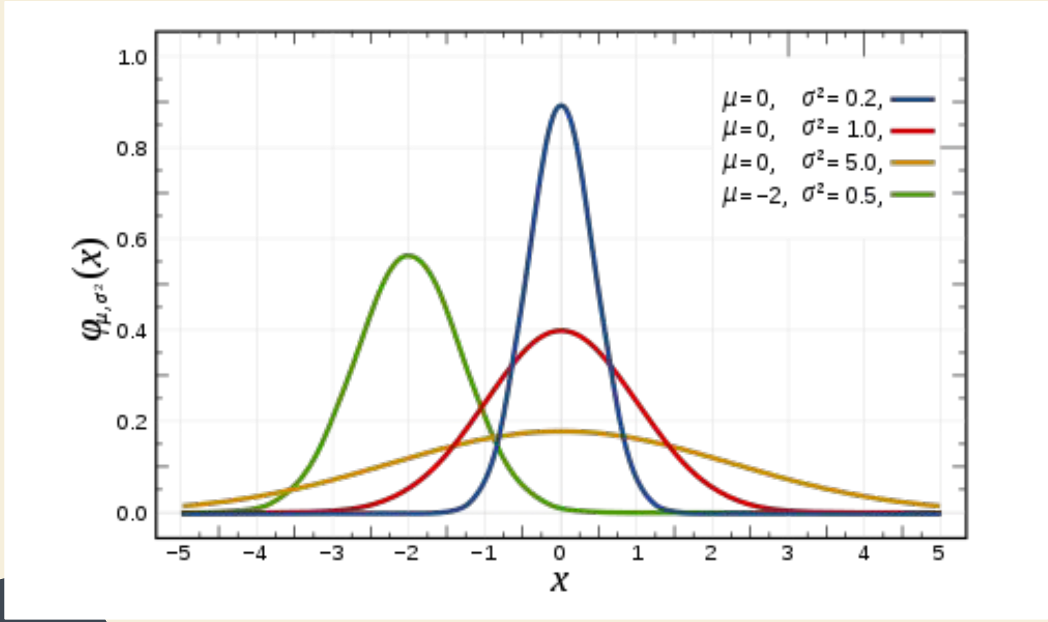
$$X \begin{cases} \boxed{\begin{array}{c} \Omega \text{ - Sample space} \\ \sigma(\Omega) \text{ - Event space} \end{array}} \\ \boxed{P \text{ - Probability Measure}} \end{cases}$$

So, a random variable isn't random nor a variable! It's just a *function*

Normal distribution (CDF)



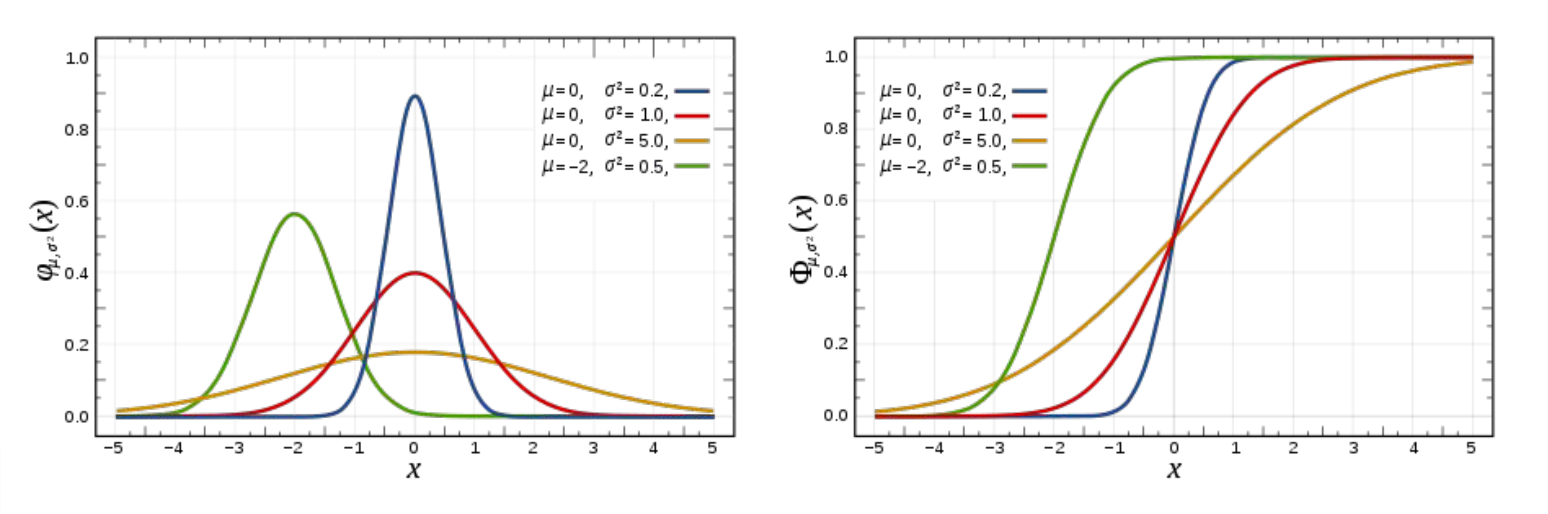
Normal distribution (PDF)



$$\frac{dF}{dx}$$

Normal distribution

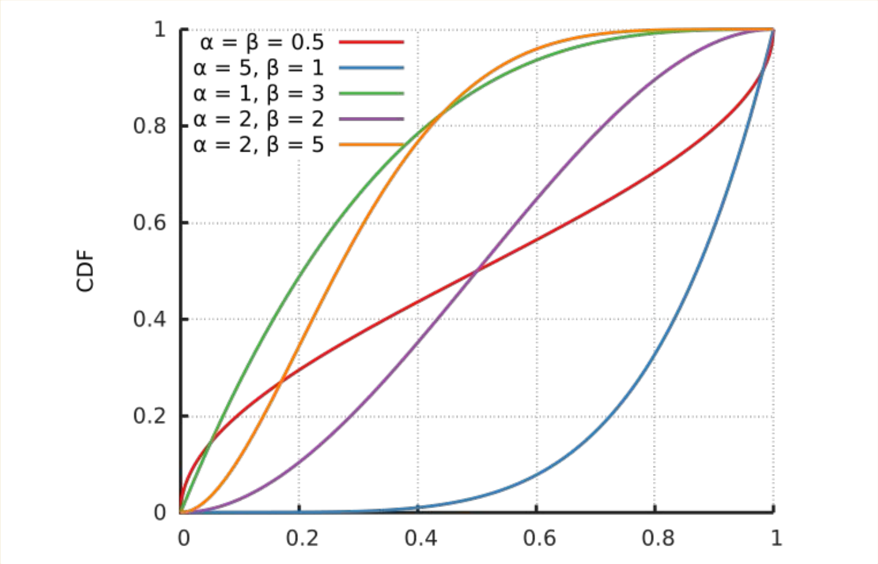
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Beta distribution

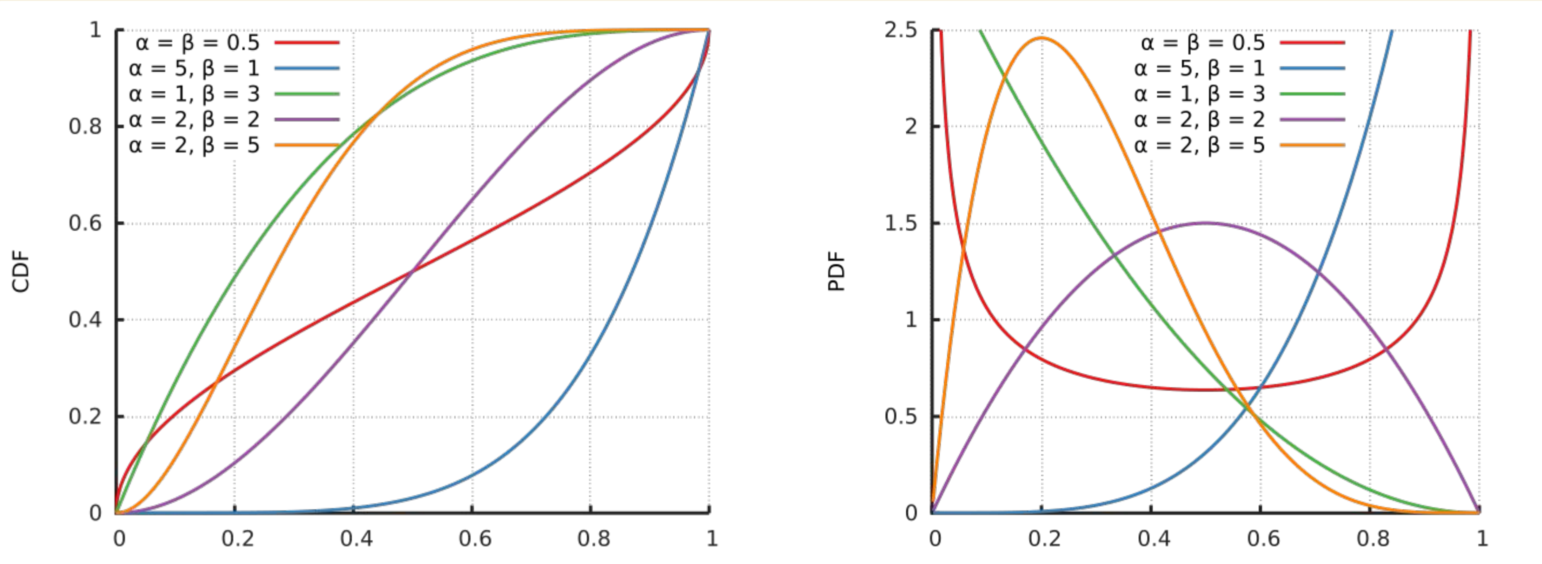
$$f(x;\alpha,\beta) = \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

Only defined on the real line in the interval [0,1]



Beta distribution

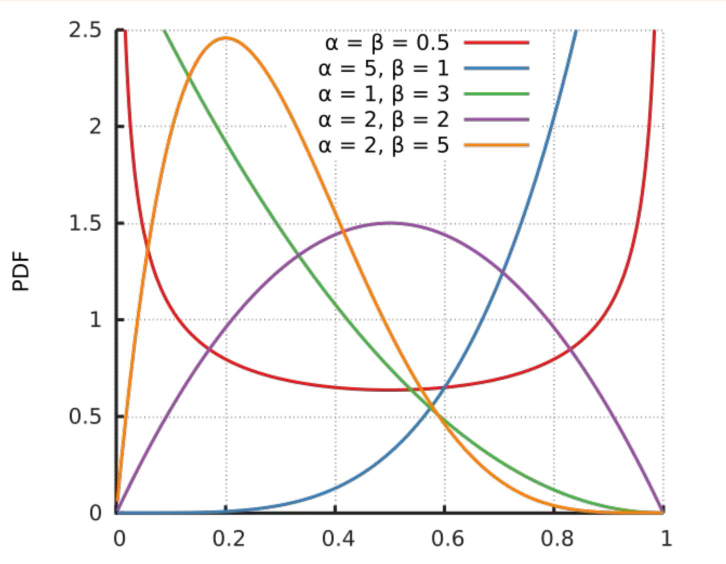
$$f(x;\alpha,\beta) = \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1}$$



Beta distribution

$$f(x; \boxed{\alpha, \beta}) = \text{constant} \cdot x^{\boxed{\alpha-1}}(1-x)^{\boxed{\beta-1}}$$

"Parameters" of a distribution are constant values that control the shape of our probability function.



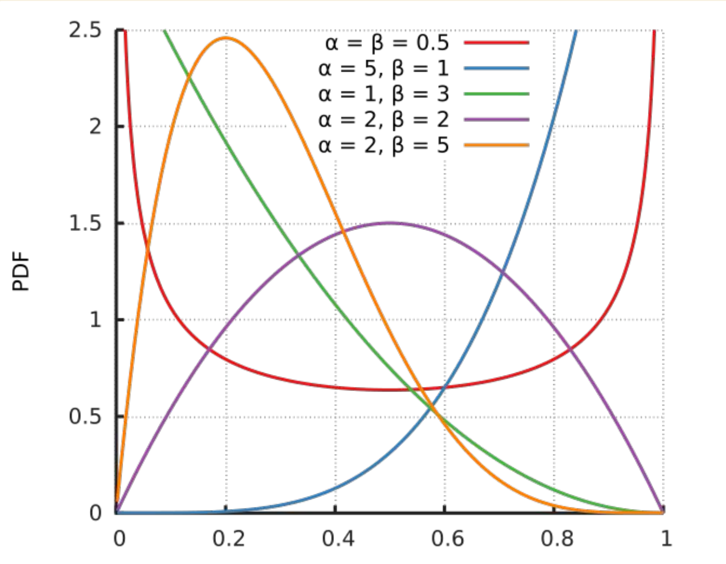
Beta distribution

$$f(x; \boxed{\alpha, \beta}) = \text{constant} \cdot x^{\boxed{\alpha-1}}(1-x)^{\boxed{\beta-1}}$$

"Parameters" of a distribution are constant values that control the shape of our probability function.

Where we view probability densities as models to learn, the parameters are the values we want to fit

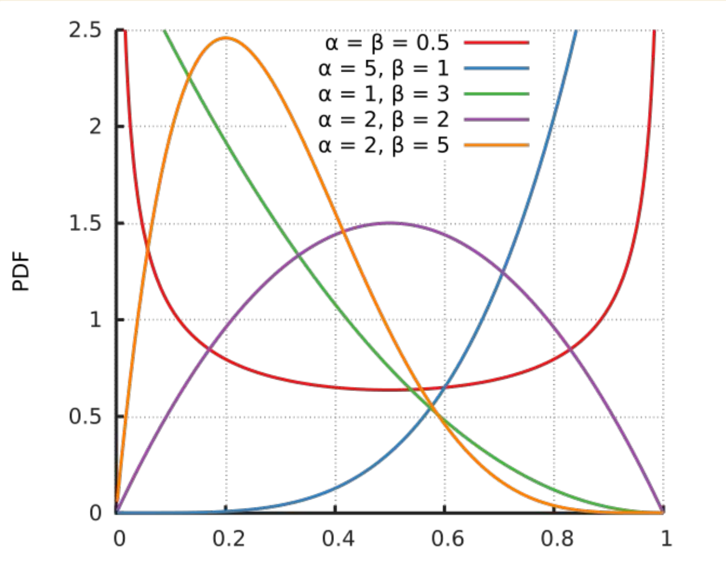
Fitting them is density estimation!



Beta distribution

$$f(x; \boxed{\alpha, \beta}) = \text{constant} \cdot x^{\boxed{\alpha - 1}}(1-x)^{\boxed{\beta - 1}}$$

Note:
The parameters of a distribution are different from the "mean" and "variance" which are indeed parameters of a normal distribution, but mean and variance are properties of a distribution, not parameters (necessarily!)



Expectation of Random Variable

$$\boxed{\mathbb{E}}_p[X]$$

- Expectation

Expectation of Random Variable

$$\boxed{\mathbb{E}}_p\left[\boxed{X}\right]$$

- Expectation
- Random variable

Expectation of Random Variable

$$\boxed{\mathbb{E}}_{\boxed{p}}[\boxed{X}]$$

Expectation of X~

- Expectation
- Random variable
- Distributed according to p

Expectation of Random Variable

$$\boxed{\mathbb{E}}_{\boxed{p}}[\boxed{X}]$$

- Expectation
- Random variable
- Distributed according to p

(I use little p when referring to specific distributions and big P when referring to the general idea of probability distributions. Feel free to use these however you like.)

Variance of a random variable

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

Variance of a random variable

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

Looking at this in the discrete case gives us a clearer intuition for what is happening here:

$$\mathrm{Var}_p[X] = \sum_{i=1}^{n} p(x_i)(x_i - \mathbb{E}[X])^2$$

thus LOTUS: $E[g(x)] = \sum p(x) g(x)$

Looking at the Exp. & Var. of distributions

$$\mathbb{E}_{p(x;\alpha,\beta)}[X] = \frac{\alpha}{(\alpha + \beta)}$$

$$\mathrm{Var}_{p(x;\alpha,\beta)}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Looking at the Exp. & Var. of distributions

$$\mathbb{E}_{p(x;\alpha,\beta)}[X] = \frac{\alpha}{(\alpha+\beta)}$$

$$\mathrm{Var}_{p(x;\alpha,\beta)}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\mathbb{E}_{p(x;\mu,\sigma)}[X] = \mu$$

$$\mathrm{Var}_{p(x;\mu,\sigma)}[X] = \sigma^2$$

Change of variables

One concept in probability that is generally very useful in machine learning (and in engineering) is the concept of change of variables in probability.

$$g(X) \qquad\qquad X \begin{cases} \Omega \text{ - Sample space} \\ \sigma(\Omega) \text{ - Event space} \\ P \text{ - Probability Measure} \end{cases}$$

A monotonic one-to-one function (we assume increasing here)

A random variable (neither random, nor variable)

Change of variables

$$g(X)$$

$$X \begin{cases} \Omega \text{ - Sample space} \\ \sigma(\Omega) \text{ - Event space} \\ P \text{ - Probability Measure} \end{cases}$$

$$Y = g(X)$$

A brand new random variable!

Change of variables

$$g(X)$$

$$X \begin{cases} \Omega \text{ - Sample space} \\ \sigma(\Omega) \text{ - Event space} \\ P \text{ - Probability Measure} \end{cases}$$

$$Y = g(X)$$

Not random, not a variable

A brand new random variable!

Change of variables

Not random, not a variable

$$Y = g(X)$$

A brand new random variable!

The distribution of this random variable is:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

Change of variables

$$Y = g(X)$$

A brand new random variable!

The distribution of this random variable is:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

How can we reason about Y?

Change of variables

$$Y = g(X)$$

We can get the probabilities of events under Y as long as we know probability of events in X

$$\begin{aligned} & P(Y \in A) \\ & = P(g(X) \in A) \\ & = P(X \in g^{-1}(A)) \end{aligned}$$

Change of variables

$$Y = g(X)$$

An example of where this property is useful:

$$p(c \leq Y \leq d) = p(c \leq g \quad (X) \leq d)$$

Change of variables

$$Y = g(X)$$

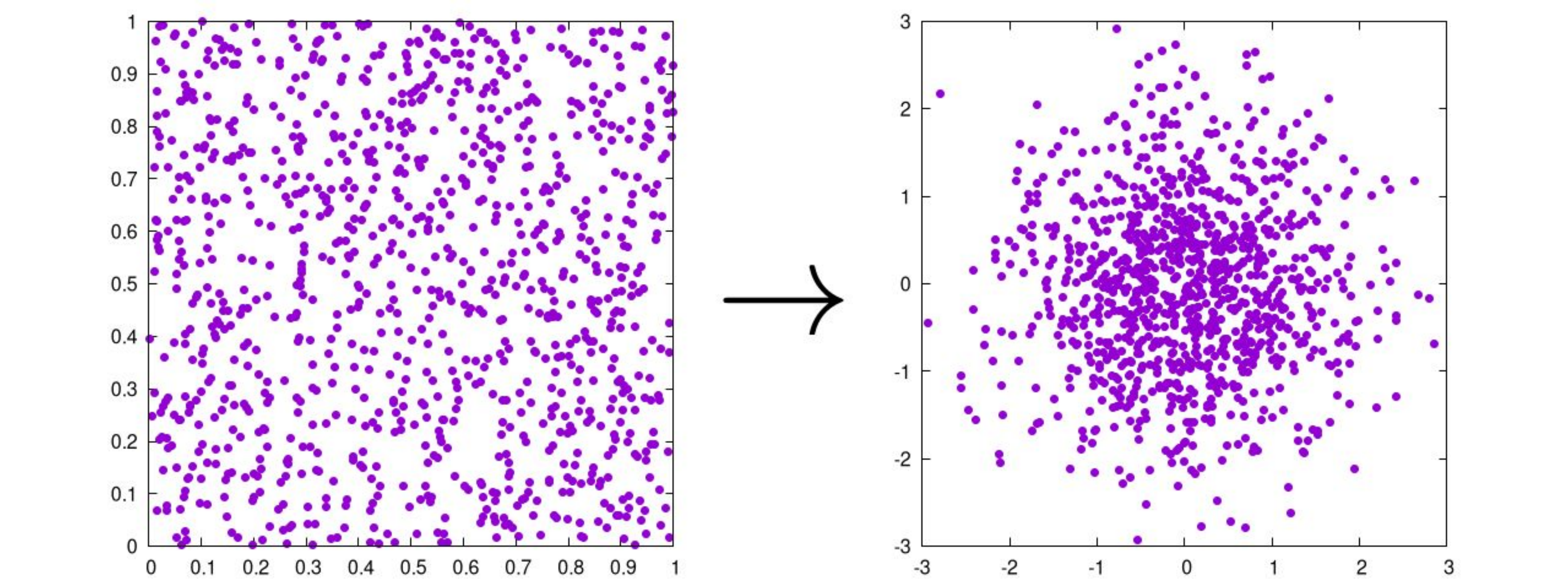
An example of where this property is useful:

$$p(c \leq Y \leq d) = p(c \leq g^{-1}(X) \leq d)$$

$$= p(g^{-1}(c) \leq g^{-1}(g(X)) \leq g^{-1}(d))$$

$$= p(a \leq X \leq b)$$

Box-Muller Transform



Law of the unconscious statistician

When we apply a change of variables, how does the expectation change?

$$X = T(Z)$$

Law of the unconscious statistician

When we apply a change of variables, how does the expectation change?

$$X = T(Z)$$

LOTUS:

$$\mathbb{E}_X[f(X)] = \mathbb{E}_Z[f(T(Z))]$$

Law of the unconscious statistician

$$X = T(Z)$$

X is a fn of Z

$$\mathbb{E}_X[f(X)] = \sum_x p_X(X = x) f(x)$$

$$\mathbb{E}_X[f(X)] = \sum_x \left(\sum_{z:T(z)=x} p_Z(Z = z) \right) f(x)$$

Law of the unconscious statistician

$$X = T(Z)$$

$$\begin{aligned}
\mathbb{E}_X[f(X)] &= \sum_x \left(\sum_{z:T(z)=x} p_Z(Z=z) \right) f(x) \\
&= \sum_z p_Z(Z=z) f(T(z)) = \mathbb{E}_Z[f(T(Z))]
\end{aligned}$$

Law of the unconscious statistician

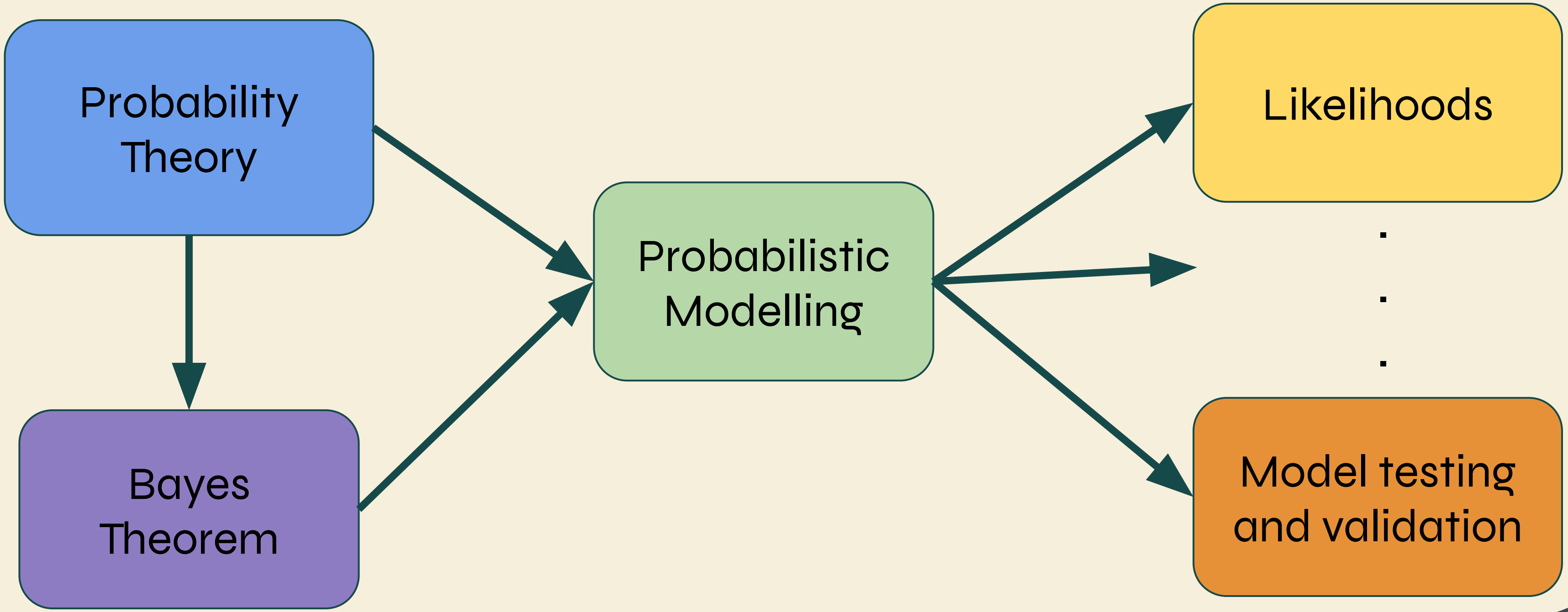
$$\mathbb{E}_X[f(X)] = \sum_x \left(\sum_{z:T(z)=x} p_Z(Z = z) \right) f(x)$$

$$= \sum_z p_Z(Z = z) f(T(z)) = \mathbb{E}_Z[f(T(Z))]$$

$$\boxed{\mathbb{E}_X[f(X)] = \mathbb{E}_Z[f(T(Z))]}$$

$$\sum_x p(X=x) f(x) = \sum_z p(Z=z) f(T(z))$$

Overview



Probability Theory
Bayes Theorem
Probabilistic Modelling
Likelihoods
.
.
.
Model testing and validation

Next lecture: Multivariate probability