### Some Essentials of Probability for Probabilistic Machine Learning

Piyush Rai  
IIT Kanpur  
Probabilistic Machine Learning (CS772A)  
Jan 4, 2016

---

### Random Variables

- A random variable (r.v.) $X$ denotes possible outcomes of an event
- Can be discrete (i.e., finite many possible outcomes) or continuous

Some examples of discrete r.v.
- A random variable $X \in \{0, 1\}$ denoting outcomes of a coin-toss
- A random variable $X \in \{1, 2, \ldots, 6\}$ denoting outcome of a dice roll

Some examples of continuous r.v.
- A random variable $X \in (0, 1)$ denoting the bias of a coin
- A random variable $X$ denoting heights of students in CS772
- A random variable $X$ denoting time to get to your hall from the department

An r.v. is associated with a probability mass function or prob. distribution

---

### Discrete Random Variables

- For a discrete r.v. $X$, $p(x)$ denotes the probability that $p(X = x)$
- $p(x)$ is called the probability mass function (PMF)

$$
\begin{align*}
  p(x) &\geq 0 \\
  p(x) &\leq 1 \\
  \sum_x p(x) &= 1
\end{align*}
$$

---

### Continuous Random Variables

- For a continuous r.v. $X$, a probability $p(X = x)$ is meaningless
- Instead we use $p(x)$ to denote the probability density function (PDF)

$$
p(x) \geq 0 \quad \text{and} \quad \int_x p(x) \, dx = 1
$$

- Probability that a cont. r.v. $X \in (x, x + \delta x)$ is $p(x)\delta x$ as $\delta x \to 0$

- Probability that $X$ lies between $(-\infty, z)$ is given by the cumulative distribution function (CDF) $P(x)$ where

$$
P(x) = p(X \leq x) = \int_{-\infty}^{x} p(x) \, dx \quad \text{and} \quad p(x) = |P'(x)|_{x=x}
$$
A word about notation..

- \( p(\cdot) \) can mean different things depending on the context
  - \( p(X) \) denotes the PMF/PDF of an r.v. \( X \)
  - \( p(X = x) \) or \( p(x) \) denotes the probability or probability density at point \( x \)
- Actual meaning should be clear from the context (but be careful)
- Exercise the same care when \( p(\cdot) \) is a specific distribution (Bernoulli, Beta, Gaussian, etc.)
- The following means drawing a sample from the distribution \( p(X) \)
  \[ x \sim p(X) \]

Joint Probability

Joint probability \( p(X, Y) \) models probability of co-occurrence of two r.v. \( X, Y \)
For discrete r.v., the joint PMF \( p(X, Y) \) is like a table (that sums to 1)

\[
\sum_x \sum_y p(X = x, Y = y) = 1
\]

\[ p(X=x, Y=y) \]

For continuous r.v., we have joint PDF \( p(X, Y) \)

\[
\int \int p(X = x, Y = y) \, dx \, dy = 1
\]

Marginal Probability

- For discrete r.v.
  \[ p(X) = \sum_y p(X, Y = y), \quad p(Y) = \sum_x p(X = x, Y) \]
- For discrete r.v., it is the sum of the PMF table along the rows/columns

\[ X \]

- For continuous r.v.
  \[ p(X) = \int_Y p(X, Y = y) \, dy, \quad p(Y) = \int_X p(X = x, Y) \, dx \]

Conditional Probability

- Meaning: Probability of one event when we know the outcome of the other
- Conditional probability \( p(X|Y) \) or \( p(Y|X) \): like taking a slice of \( p(X, Y) \)
- For a discrete distribution:

\[ Y \]

- For a continuous distribution\(^1\):

\(^1\)Pinned reference: Computer vision: models, learning and inference [Oksa, Pietikäinen]
**Some Basic Rules**

- **Sum rule:** Gives the marginal probability
  - For discrete r.v.: \( p(X) = \sum_Y p(X, Y) \)
  - For continuous r.v.: \( p(X) = \int_Y p(X, Y) dY \)
- **Product rule:** \( p(X, Y) = p(Y|X) p(X) = p(X|Y) p(Y) \)
- **Bayes rule:** Gives conditional probability
  \[
  p(Y|X) = \frac{p(X|Y) p(Y)}{p(X)}
  \]
  - For discrete r.v.: \( p(Y|X) = \frac{p(X|Y) p(Y)}{p(X)} \)
  - For continuous r.v.: \( p(Y|X) = \frac{p(X|Y) p(Y)}{p(X)} \)
- Bayes rule is also central to parameter estimation (more on this later)
- Also remember the chain rule
  \[
  p(X_1, X_2, \ldots, X_n) = p(X_1)p(X_2|X_1)\cdots p(X_n|X_1, \ldots, X_{n-1})
  \]

**Independence**

- \( X \) and \( Y \) are independent \( (X \perp Y) \) when one tells nothing about the other
  \[
  p(X|Y) = p(X) \quad p(Y|X) = p(Y) \quad p(X,Y) = p(X)p(Y)
  \]
- \( X \perp Y \) is also called marginal independence
- **Conditional independence** \( (X \perp Y|Z) \): independence when another event \( Z \) is observed
  \[
  p(X, Y|Z) = p(X|Z)p(Y|Z)
  \]

**Expectation**

- **Expectation or mean \( \mu \) of an r.v. with PMF/PDF \( p(X) \)**
  \[
  E[X] = \sum_x x p(x) \quad \text{(for discrete distributions)}
  
  E[X] = \int_x x p(x) dx \quad \text{(for continuous distributions)}
  \]
- **Note:** The definition applies to functions of r.v. too (e.g., \( E[f(x)] \))
- **Linearity of expectation** (very important/useful property)
  \[
  E[\alpha f(x) + \beta g(x)] = \alpha E[f(x)] + \beta E[g(x)]
  \]

**Variance and Covariance**

- **Variance** \( \sigma^2 \) (or “spread” around mean) of an r.v. with PMF/PDF \( p(X) \)
  \[
  \text{var}[X] = E[(X - \mu)^2] = E[X^2] - \mu^2
  \]
- **Standard deviation:** \( \text{std}[X] = \sqrt{\text{var}[X]} = \sigma \)
- **Note:** The definition applies to functions of r.v. too (e.g., \( \text{var}[f(X)] \))
- For r.v. \( x \) and \( y \), the **covariance** is defined by
  \[
  \text{cov}[x, y] = E_{x,y}[(x - E[x])(y - E[y])] = E_{x,y}[xy] - E[x]E[y]
  \]
- For vector r.v. \( x \) and \( y \), the covariance matrix is defined as
  \[
  \text{cov}[x, y] = E_{x,y}[(x - E[x])(y - E[y])^T] = E_{x,y}[xy^T] - E[x]E[y^T]
  \]
- Cov. of components of a vector r.v. \( x \) with each other: \( \text{cov}[x] = \text{cov}[x, x] \)
Transformation of Random Variables

Suppose $y = f(x) = Ax + b$ be a linear function of an r.v. $x$
Suppose $E[x] = \mu$ and $\text{cov}[x] = \Sigma$

- Expectation of $y$: $E[y] = E[Ax + b] = A\mu + b$
- Covariance of $y$: $\text{cov}[y] = \text{cov}[Ax + b] = A\Sigma A^T$

Likewise if $y = f(x) = a^T x + b$ is a scalar-valued linear function of an r.v. $x$:

- $E[y] = E[a^T x + b] = a^T \mu + b$
- $\text{var}[y] = \text{var}[a^T x + b] = a^T \Sigma a$

Common Probability Distributions

**Important:** We will use these extensively to model data as well as parameters
Some discrete distributions and what they can model:
- **Bernoulli:** Binary numbers, e.g., outcome (head/tail, 0/1) of a coin toss
- **Binomial:** Bounded non-negative integers, e.g., # of heads in $n$ coin tosses
- **Multinomial:** One of $K (>2)$ possibilities, e.g., outcome of a dice roll
- **Poisson:** Non-negative integers, e.g., # of words in a document
- .. and many others

Some continuous distributions and what they can model:
- **Uniform:** numbers defined over a fixed range
- **Beta:** numbers between 0 and 1, e.g., probability of head for a biased coin
- **Gamma:** Positive unbounded real numbers
- **Dirichlet:** vectors that sum of 1 (fraction of data points in different clusters)
- **Gaussian:** real-valued numbers or real-valued vectors
- .. and many others

Discrete Distributions

**Bernoulli Distribution**

- Distribution over a binary r.v. $x \in \{0, 1\}$, like a coin-toss outcome
- Defined by a probability parameter $p \in (0, 1)$

\[ P(x = 1) = p \]

- Distribution defined as: $\text{Bernoulli}(x; p) = p^x (1 - p)^{1-x}$

- Mean: $E[x] = p$
- Variance: $\text{var}[x] = p(1 - p)$
**Binomial Distribution**

- Distribution over number of successes $m$ (an r.v.) in a number of trials $N$.
- Defined by two parameters: total number of trials $N$ and probability of each success $p \in (0, 1)$.
- Can think of Binomial as multiple independent Bernoulli trials.
- Distribution defined as
  \[
  \text{Binomial}(m; N, p) = \binom{N}{m} p^m (1-p)^{N-m}
  \]

- Mean: $\mathbb{E}[m] = Np$
- Variance: $\text{var}[m] = Np(1-p)$

**Multinoulli Distribution**

- Also known as the categorical distribution (models categorical variables).
- Think of a random assignment of an item to one of $K$ bins - a $K$ dim. binary r.v. $x$ with single 1 (i.e., $\sum_{k=1}^{K} x_k = 1$). Modeled by a multinoulli
  \[
  \begin{bmatrix}
  0 & 0 & \ldots & 0 & 1 & 0 & 0
  \end{bmatrix}
  \]
- Let vector $p = [p_1, p_2, \ldots, p_K]$ define the probability of going to each bin.
  - $p_k \in (0, 1)$ is the probability that $x_k = 1$ (assigned to bin $k$).
  - $\sum_{k=1}^{K} p_k = 1$
- The multinoulli is defined as: $\text{Multinoulli}(x; p) = \prod_{k=1}^{K} p_k^{x_k}$
- Mean: $\mathbb{E}[x_k] = p_k$
- Variance: $\text{var}[x_k] = p_k (1-p_k)$

**Multinomial Distribution**

- Think of repeating the Multinoulli $N$ times.
- Like distributing $N$ items to $K$ bins. Suppose $x_k$ is count in bin $k$.
  \[
  0 \leq x_k \leq N \quad \forall \ k = 1, \ldots, K, \quad \sum_{k=1}^{K} x_k = N
  \]
- Assume probability of going to each bin: $p = [p_1, p_2, \ldots, p_K]$.
- Multinomial models the bin allocations via a discrete vector $x$ of size $K$.
- Distribution defined as
  \[
  \text{Multinomial}(x; N, p) = \binom{N}{x_1, x_2, \ldots, x_K} \prod_{k=1}^{K} p_k^{x_k}
  \]
- Mean: $\mathbb{E}[x_k] = Np_k$
- Variance: $\text{var}[x_k] = Np_k (1-p_k)$
- Note: For $N = 1$, multinomial is the same as multinoulli.

**Multinoulli/Multinomial: Pictorially**

![Multinomial Distribution Diagram](image_url)
**Poisson Distribution**

- Used to model a non-negative integer (count) r.v. $k$
- Examples: number of words in a document, number of events in a fixed interval of time, etc.
- Defined by a positive rate parameter $\lambda$
- Distribution defined as
  
  $\text{Poisson}(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, 1, 2, \ldots$

- Mean: $\mathbb{E}[k] = \lambda$
- Variance: $\text{var}[k] = \lambda$

**Continuous Distributions**

**Uniform Distribution**

- Models a continuous r.v. $x$ distributed uniformly over a finite interval $[a, b]$
  
  $\text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b - a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$

- Mean: $\mathbb{E}[x] = \frac{(a+b)}{2}$
- Variance: $\text{var}[x] = \frac{(b-a)^2}{12}$

**Beta Distribution**

- Used to model an r.v. $p$ between 0 and 1 (e.g., a probability)
- Defined by two shape parameters $\alpha$ and $\beta$
  
  $\text{Beta}(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

- Mean: $\mathbb{E}[p] = \frac{\alpha}{\alpha + \beta}$
- Variance: $\text{var}[p] = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$
- Often used to model the probability parameter of a Bernoulli or Binomial (also conjugate to these distributions)
Gamma Distribution

- Used to model positive real-valued r.v. x
- Defined by a shape parameter $k$ and a scale parameter $\theta$

\[ \text{Gamma}(x; k, \theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k B(k)} \]

- Mean: $E[x] = k\theta$
- Variance: $\text{var}[x] = k\theta^2$
- Often used to model the rate parameter of Poisson or exponential distribution, or to model the inverse variance of a Gaussian

Dirichlet Distribution

- Used to model non-negative r.v. vectors $p = [p_1, \ldots, p_K]$ that sum to 1
  \[ 0 \leq p_k \leq 1, \quad \forall k = 1, \ldots, K \quad \text{and} \quad \sum_{k=1}^{K} p_k = 1 \]
- Equivalent to a distribution over the $K-1$ dimensional simplex
- Defined by a $K$ size vector $\alpha = [\alpha_1, \ldots, \alpha_K]$ of positive reals
- Distribution defined as
  \[ \text{Dirichlet}(p; \alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \]
- Often used to model parameters of Multinomial/Multinomial
- Dirichlet is conjugate to Multinomial/Multinomial
- Note: Dirichlet can be seen as a generalization of the Beta distribution. Normalizing a bunch of Gamma r.v.'s gives an r.v. that is Dirichlet distributed.

Now comes the Gaussian (Normal) distribution..
**Univariate Gaussian Distribution**
- Distribution over real-valued scalar r.v. \( x \)
- Defined by a scalar mean \( \mu \) and a scalar variance \( \sigma^2 \)
- Distribution defined as
  \[
  N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
  \]
- Mean: \( \mathbb{E}[x] = \mu \)
- Variance: \( \text{var}[x] = \sigma^2 \)
- Precision (inverse variance) \( \beta = 1/\sigma^2 \)

**Multivariate Gaussian Distribution**
- Distribution over a multivariate r.v. vector \( x \in \mathbb{R}^D \) of real numbers
- Defined by a mean vector \( \mu \in \mathbb{R}^D \) and a \( D \times D \) covariance matrix \( \Sigma \)
  \[
  N(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}
  \]
- The covariance matrix \( \Sigma \) must be symmetric and positive definite
  - All eigenvalues are positive
  - \( z^T \Sigma z > 0 \) for any real vector \( z \)
- Often we parameterize a multivariate Gaussian using the inverse of the covariance matrix, i.e., the precision matrix \( \Lambda = \Sigma^{-1} \)

**Multivariate Gaussian: The Covariance Matrix**
- The covariance matrix can be spherical, diagonal, or full

### Spherical Covariances
- \( \Sigma = \sigma^2 I \)

### Diagonal Covariances
- \( \Sigma = \text{diag} \{ \sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2 \} \)

### Full Covariances
- \( \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12} \sigma_1 \sigma_2 & \cdots & \rho_{1D} \sigma_1 \sigma_D \\ \rho_{21} \sigma_1 \sigma_2 & \sigma_2^2 & \cdots & \rho_{2D} \sigma_2 \sigma_D \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{D1} \sigma_1 \sigma_D & \rho_{D2} \sigma_2 \sigma_D & \cdots & \sigma_D^2 \end{pmatrix} \)

---

Some nice properties of the Gaussian distribution.
Multivariate Gaussian: Marginals and Conditionals

- Given jointly Gaussian distribution \( \mathcal{N}(x; \mu, \Sigma) \) with \( \Lambda = \Sigma^{-1} \) with

\[
\begin{align*}
    x &= \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix}, \\
    \mu &= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \\
    \Sigma &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \\
    \Lambda &= \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}
\end{align*}
\]

- The marginal distribution is simply

\[
p(x_i) = \mathcal{N}(x_i; \mu_i, \Sigma_{ii})
\]

- The conditional distribution is given by

\[
p(x_i|x_j) = \mathcal{N}(x_i; \mu_{ij}, \Lambda_{ij})
\]

\[
\mu_{ij} = \mu_i - \Lambda_{ij}\Lambda_{jj}(x_j - \mu_j)
\]

Thus marginals and conditionals of Gaussians are Gaussians.

Multivariate Gaussian: Marginals and Conditionals

- Given the conditional and marginal of r.v. being conditioned on

\[
\begin{align*}
p(y|x) &= \mathcal{N}(y; A\mu + b, L^{-1}) \\
p(x) &= \mathcal{N}(x; \mu, \Lambda^{-1})
\end{align*}
\]

- Marginal and “reverse” conditional are given by

\[
\begin{align*}
p(y) &= \mathcal{N}(y; A\mu + b, L^{-1} + A\Lambda^{-1}A^T) \\
p(x|y) &= \mathcal{N}(x; \Sigma(A^TL(y - b) + A\mu), \Sigma)
\end{align*}
\]

where \( \Sigma = (\Lambda + A^T LA)^{-1} \)

- Very useful property for probabilistic models with Gaussian likelihoods and/or priors. Also very handy for computing marginal likelihoods.

Gaussians: Product of Gaussians

- Pointwise multiplication of two Gaussians is another (unnormalized) Gaussian

\[
\begin{align*}
    \mathcal{N}(x; \mu, \Sigma) \mathcal{N}(x; \nu, P) &= \frac{1}{Z} \mathcal{N}(x; \omega, T), \\
    \text{where} \\
    T &= (\Sigma^{-1} + P^{-1})^{-1} \\
    \omega &= T(\Sigma^{-1}\mu + P^{-1}\nu) \\
    Z^{-1} &= \mathcal{N}(\mu; \nu, \Sigma + P) = \mathcal{N}(\nu; \mu, \Sigma + P)
\end{align*}
\]

Multivariate Gaussian: Affine Transforms

- Given a \( x \in \mathbb{R}^d \) with a multivariate Gaussian distribution

\[
\mathcal{N}(x; \mu, \Sigma)
\]

- Consider an affine transform of \( x \) into \( \mathbb{R}^D \)

\[
y = Ax + b
\]

where \( A \) is \( D \times d \) and \( b \in \mathbb{R}^D \)

- \( y \in \mathbb{R}^D \) will have a multivariate Gaussian distribution

\[
\mathcal{N}(y; A\mu + b, A\Sigma A^T)
\]
Exponential Family

- An exponential family distribution is defined as
  \[ p(x; \theta) = h(x) e^{\theta^T T(x) - A(\theta)} \]
- \( \theta \) is called the parameter of the family
- \( h(x), \eta(\theta), T(x), \text{ and } A(\theta) \) are known functions
- \( p(\cdot) \) depends on \( x \) only through \( T(x) \)
- \( T(x) \) is called the sufficient statistics: summarizes the entire \( p(x; \theta) \)
- Exponential family is the only family for which conjugate priors exist (often also in the exponential family)
- Many other nice properties (especially useful in Bayesian inference)

Many well-known distributions (Bernoulli, Binomial, categorical, beta, gamma, Gaussian, etc.) are exponential family distributions

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

Binomial as Exponential Family

- Recall the exponential family distribution
  \[ p(x; \theta) = h(x) e^{\theta^T T(x) - A(\theta)} \]
- Binomial in the usual form:
  \[ \text{Binomial}(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \]
- Can re-express it as
  \[ \binom{n}{x} e^{\eta(\theta) + n \log(1 - p)} \]
  \[ h(x) = \binom{n}{x} \]
  \[ \eta(\theta) = \log \left( \frac{p}{1 - p} \right) \]
  \[ T(x) = x \]
  \[ A(\theta) = -n \log(1 - p) \]

Gaussian as Exponential Family

- Recall the exponential family distribution
  \[ p(x; \theta) = h(x) e^{\theta^T T(x) - A(\theta)} \]
- Gaussian in the usual form:
  \[ \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \]
- Can re-express it as \( p(x; \theta) = h(x) e^{\theta(\mu)} T(x) - A(\theta) \) where
  - \( h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \)
  - \( \eta(\theta) = \left( \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) \)
  - \( T(x) = (x, x^2) \)
  - \( A(\theta) = \frac{\mu^2}{2\sigma^2} + \log \sigma \)

Conjugate Priors

- Given a distribution \( p(x|\theta) \)
- We say \( p(\theta) \) is conjugate to \( p(x|\theta) \) if
  \[ p(\theta|x) \propto p(x|\theta)p(\theta) \]
  has the same form as \( p(\theta) \)
- Many pairs of distributions are conjugate to each other, e.g.,
  - Gaussian-Gaussian
  - Bernoulli-Beta
  - Poisson-Gamma
  ... and many others
- More on this in the next class.
Next class: Parameter estimation in probabilistic models