CS761 Spring 2013 Advanced Machine Learning

Probability Background

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

1 Probability Measure

A *sample space* Ω is the set of all possible outcomes. Elements $\omega \in \Omega$ are called *sample outcomes*, while subsets $A \subseteq \Omega$ are called *events*. For example, for a die roll, $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\omega = 5$ is an outcome, $A_1 = \{5\}$ is the event that the outcome is 5, and $A_2 = \{1, 3, 5\}$ is the event that the outcome is odd. Ideally, one would like to be able to assign probabilities to all As. This is trivial for finite Ω. However, when $\Omega = \mathbb{R}$ strange things happen: it is no longer possible to consistently assign probabilities to all subsets of $\mathbb{R}$.

★ *The problem is not singleton sets such as $A = \{\pi\}$, which happily has probability 0 (we will define it formally). This happens because of the existence of* non-measurable sets*, whose construction is non-trivial. See for example* `http://www.math.kth.se/matstat/gru/godis/nonmeas.pdf`

Instead, we will restrict ourselves to only some of the events. A σ-algebra $\mathcal{B}$ is a set of Ω subsets satisfying:

1. $\emptyset \in \mathcal{B}$
2. if $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$ (complementarity)
3. if $A_1, A_2, \ldots \in \mathcal{B}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ (countable unions)

The sets in $\mathcal{B}$ are called *measurable*. The pair $(\Omega, \mathcal{B})$ is called a *measurable space*. For $\Omega = \mathbb{R}$, we take the smallest σ-algebra that contains all the open subsets and call it the *Borel σ-algebra*.

★ *It turns out the Borel σ-algebra can be defined alternatively as the smallest σ-algebra that contains all the* closed *subsets. Can you see why? It also follows that singleton sets $\{x\}$ where $x \in \mathbb{R}$ is in the Borel σ-algebra.*

A *measure* is a function $P : \mathcal{B} \mapsto \mathbb{R}$ satisfying:

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
2. If $A_1, A_2, \ldots$ are disjoint then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Note these imply $P(\emptyset) = 0$. The triple $(\Omega, \mathcal{B}, P)$ is called a *measure space*. The *Lebesgue measure* is a uniform measure over the Borel σ-algebra with the usual meaning of length, area or volume, depending on dimensionality. For example, for $\mathbb{R}$ the Lebesgue measure of the interval (a, b) is $b - a$.

A *probability measure* is a measure satisfying additionally the normalization constraint:

3 $P(\Omega) = 1$.

Such a triple $(\Omega, \mathcal{B}, P)$ is called a *probability space*.

★ *When Ω is finite, $P(\{\omega\})$ has the intuitive meaning: the chance that the outcome is ω. When $\Omega = \mathbb{R}$, $P((a, b))$ is the "probability mass" assigned to the interval (a, b).*

2 Random Variables

Let $(\Omega, \mathcal{B}, P)$ be a probability space. Let $(\mathbb{R}, \mathcal{R})$ be the usual measurable space of reals and its Borel σ-algebra. A *random variable* is a function $X : \Omega \mapsto \mathbb{R}$ such that the preimage of any set $A \in \mathcal{R}$ is measurable in $\mathcal{B}$: $X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{B}$. This allows us to define the following (the first P is the new definition, while the 2nd and 3rd Ps are the already-defined probability measure on $\mathcal{B}$):

$$P(X \in A) = P(X^{-1}(A)) = P(\{\omega : X(\omega) \in A\}) \tag{1}$$

$$P(X = x) = P(X^{-1}(x)) = P(\{\omega : X(\omega) = x\}) \tag{2}$$

★ *A random variable is a function. Intuitively, the world generates random outcomes ω, while a random variable deterministically translates them into numbers.*

Example 1 *Let $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ be the unit disk. Consider drawing a point at random from Ω. The outcome is of the form $\omega = (x, y)$. Some example random variables are $X(\omega) = x, Y(\omega) = y, Z(\omega) = xy$, and $W(\omega) = \sqrt{x^2 + y^2}$.*

Example 2 *Let $X = \omega$ be a uniform random variable (to be defined later) with the sample space $\Omega = [0, 1]$. A sequence of different random variables $\{X_n\}_{n=1}^{\infty}$ can be defined as follows, where $1\{z\} = 1$ if z is true, and 0 otherwise:*

$$X_1(\omega) = \omega + 1\{\omega \in [0, 1]\} \tag{3}$$

$$X_2(\omega) = \omega + 1\{\omega \in [0, \frac{1}{2}]\} \tag{4}$$

$$X_3(\omega) = \omega + 1\{\omega \in [\frac{1}{2}, 1]\} \tag{5}$$

$$X_4(\omega) = \omega + 1\{\omega \in [0, \frac{1}{3}]\} \tag{6}$$

$$X_5(\omega) = \omega + 1\{\omega \in [\frac{1}{3}, \frac{2}{3}]\} \tag{7}$$

$$X_6(\omega) = \omega + 1\{\omega \in [\frac{2}{3}, 1]\} \tag{8}$$

$$\ldots$$

Given a random variable X, the *cumulative distribution function (CDF)* is the function $F_X : \mathbb{R} \mapsto [0, 1]$

$$F_X(x) = P(X \leq x) = P(X \in (-\infty, x]) = P(\{\omega : X(\omega) \in (-\infty, x]\}). \tag{9}$$

A random variable X is *discrete* if it takes countably many values. We define the probability mass function $f_X(x) = P(X = x)$.

A random variable X is *continuous* if there exists a function f_X such that

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. for every $a \leq b$, $P(X \in [a, b]) = \int_a^b f_X(x)dx$.

The function f_X is called the *probability density function (PDF)* of X. CDF and PDF are related by

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{10}$$

$$f_X(x) = F_X'(x) \text{ at all points } x \text{ at which } F_X \text{ is differentiable.} \tag{11}$$

★ *It is certainly possible for the PDF of a continuous random variable to be larger than one: $f_X(x) \gg 1$. In fact, the PDF can be unbounded as in $f(x) = \frac{2}{3}x^{-1/3}$ for $x \in (0, 1)$ and $f(x) = 0$ otherwise. However, $P(x) = 0$ for all x. Recall that P and f_X is related by integration over an interval. We will often use p instead of f_X to denote a PDF later in class.*

3 Some Random Variables

3.1 Discrete

Dirac or point mass distribution $X \sim \delta_a$ if $P(X = a) = 1$ with CDF $F(x) = 0$ if $x < a$ and 1 if $x \geq a$.

Binomial. A random variable X has a binomial distribution with parameters n (number of trials) and p (head probability) $X \sim \text{Binomial}(n, p)$ with probability mass function

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

If $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$ then $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$. Think this as merging two coin flip experiments on the same coin.

★ *The binomial distribution is used to model test set error of a classifier. Assuming a classifier's true error rate is p (with respect to the unknown underlying joint distribution – we will make it precise later in class), then on a test set of size n the number of misclassified items follow* $\text{Binomial}(n, p)$.

Bernoulli. Binomial with $n = 1$.

Multinomial. The d-dimensional version of binomial. The parameter $p = (p_1, \ldots, p_d)^\top$ is now the probabilities of a d-sided die, and $x = (x_1, \ldots, x_d)^\top$ is the counts of each face.

$$f(x) = \begin{cases} \binom{n}{x_1, \ldots, x_d} \prod_{k=1}^d p_k^{x_k} & \text{if } \sum_{k=1}^d x_k = n \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

★ *The multinomial distribution is typically used with the bag-of-word representation of text documents.*

Poisson. $X \sim \text{Poisson}(\lambda)$ if $f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$. If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

★ *This is a distribution on unbounded counts with a probability mass function"hump" (mode – not the mean – at $\lceil \lambda \rceil - 1$). It can be used, for example, to model the length of a document.*

Geometric. $X \sim \text{Geom}(p)$ if $f(x) = p(1-p)^{x-1}$ for $x = 1, 2, \ldots$

★ *This is another distribution on unbounded counts. Its probability mass function has no "hump" (mode – not the mean – at 1) and decreases monotonically. Think of it as a "stick breaking" procedure: start with a stick of length 1. Each day you take away p fraction of the remaining stick. Then $f(x)$ is the length you get on day x. We will see more interesting stick breaking in Bayesian nonparametrics.*

3.2 Continuous

Gaussian (also called Normal distribution). $X \sim N(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ (the mean) and σ^2 (the variance) if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{14}$$

The square root of variance $\sigma > 0$ is called the standard deviation. If $\mu = 0, \sigma = 1$, X has a *standard normal distribution.* In this case, X is usually written as Z. Some useful properties:

- (Scaling) If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$
- (Independent sum) If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, then $\sum_i X_i \sim N\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$

★ *Note there is concentration of measure in the independent sum: the "spread" or stddev grows as $\sqrt{n}$, not n.*

★ *This is one PDF you need to pay close attention to. The CDF of a standard normal is usually written as $\Phi(z)$, which has no closed-form expression. However, it qualitatively resembles a sigmoid function and is often used in translating unbounded real "margin" values into probabilities of binary classes. You might recognize the RBF kernel as an unnormalized version of the Gaussian PDF. The parameter σ^2 or confusingly σ or a scaled version of either is called the bandwidth.*

χ^2 distribution. If $Z_1, \ldots, Z_k$ are independent standard normal random variables, then $Y = \sum_i^k Z_i^2$ has a χ^2 distribution with k degrees of freedom. If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, then $\sum_i \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ has a χ^2 distribution with k degrees of freedom. The PDF for the χ^2 distribution with k degrees of freedom is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \; x > 0. \tag{15}$$

Multivariate Gaussian. Let $x, \mu \in \mathbb{R}^d$, $\Sigma \in S_+^d$ a symmetric, positive definite matrix of size $d \times d$. Then $X \sim N(\mu, \Sigma)$ with PDF

$$f(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \tag{16}$$

Here, μ is the mean vector, Σ is the covariance matrix, $|\Sigma|$ its determinant, and Σ^{-1} its inverse (exists because Σ is positive definite).

If we have two (groups of) variables that are jointly Gaussian:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix}\right) \tag{17}$$

then we have:

- (Marginal) $x \sim N(\mu_x, A)$
- (Conditional) $y|x \sim N(\mu_y + C^\top A^{-1}(x - \mu_x), B - C^\top A^{-1} C)$

★ *You want to pay attention to Multivariate Gaussian too. It is used frequently in mixture models (Mixture of Gaussian), and as building blocks for Gaussian Processes. The conditional is useful for inference.*

Exponential. $X \sim \text{Exp}(\beta)$ with parameter $\beta > 0$, if $f(x) = \frac{1}{\beta} e^{-x/\beta}$.

★ *Not to be confused with the exponential family (more later).*

Gamma. The *Gamma function* (not distribution) is defined as $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$ with $\alpha > 0$. The Gamma function is an extension to the factorial function: $\Gamma(n) = (n-1)!$ when n is a positive integer. It also satisfies $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$. X has a *Gamma distribution* with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, denoted by $X \sim \text{Gamma}(\alpha, \beta)$, if

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \; x > 0. \tag{18}$$

$\text{Gamma}(1, \beta)$ is the same as $\text{Exp}(\beta)$.

Beta. $X \sim \text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$, if

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \; x \in (0, 1). \tag{19}$$

$\text{Beta}(1, 1)$ is uniform in $[0, 1]$. $\text{Beta}(\alpha < 1, \beta < 1)$ has a U-shape. $\text{Beta}(\alpha > 1, \beta > 1)$ is unimodal with mean $\alpha/(\alpha + \beta)$ and mode $(\alpha - 1)/(\alpha + \beta - 2)$.

★ *Beta distribution has a special status in machine learning because it is the conjugate prior of the binomial and Bernoulli distributions. A draw from a beta distribution can be thought of as generating a (biased) coin. A draw from the corresponding Bernoulli distribution can be thought of as a flip of that coin.*

Dirichlet. The Dirichlet distribution is the multivariate version of the beta distribution. $X \sim \text{Dir}(\alpha_1, \ldots, \alpha_d)$ with parameters $\alpha_i > 0$, if

$$f(x) = \frac{\Gamma(\sum_i^d \alpha_i)}{\prod_i^d \Gamma(\alpha_i)} \prod_i^d x_i^{\alpha_i - 1}, \tag{20}$$

where $x = (x_1, \ldots, x_d)$ with $x_i > 0, \sum_i^d x_i = 1$. This support is called the open $(d-1)$ dimensional simplex.
★ *The Dirichlet distribution is important for machine learning because it is the conjugate prior for multinomial distributions. The analogy here is that of a dice factory (Dirichlet) and die rolls (multinomial). It is used extensively for modeling bag-of-word documents. We will also see it in Dirichlet Processes.*

t **(or Student's t) Distribution** $X \sim t_\nu$ has a t distribution with ν degrees of freedom, if

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} \tag{21}$$

It is similar to a Gaussian distribution, but with heavier tails (more likely to produce extreme values). When $\nu = \infty$, it becomes a standard Gaussian distribution.
★ *"Student" is the pen name of William Gosset who worked for Guinness, because the brewery did not want its employees to publish any scientific papers for the fear of leaking trade secret (sounds familiar?).*

Cauchy. The Cauchy distribution is a special case of t-distribution with $\nu = 1$. The PDF is

$$f(x) = \frac{1}{\pi\gamma\left(1 + \left(\frac{x - x_0}{\gamma}\right)^2\right)} \tag{22}$$

where x_0 is the location parameter for the mode, and γ is the scale parameter. It is notable for the lack of mean: $\mathbb{E}(X)$ does not exist because $\int_x |x| dF_X(x) = \infty$. Similarly, it has no variance or higher moments. However, the mode and median are both x_0.
★ *This is a very heavy-tailed distribution compared to Gaussian, so much so that draws from Cauchy will occasionally produce very large values and the sample mean never settles.*

4 Convergence of Random Variables

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables, and X be another random variable.
★ *Think of X_n as the classifier trained on a training set of size n. It is a random variable because the training set is random (an iid sample of the underlying unknown joint distribution P_{XY}). Think of X as the "true" classifier (well, it is a fixed quantity, not random, but that's fine). We are interested in how the classifiers behave as we have more and more training data: Do they get closer to X? What do we mean by "close?" These questions are made precise by the different types of convergence. Study of such topics are called large sample theory or asymptotic theory.*

X_n **converges to X in distribution**, written $X_n \rightsquigarrow X$, if

$$\lim_{n\to\infty} F_{X_n}(t) = F(t) \tag{23}$$

at all t where F is continuous. Here, F_{X_n} is the CDF of X_n, and F is the CDF of X. We expect to see the next outcome in a sequence of random experiments becoming better and better modeled by the probability distribution of X. In other words, the probability for X_n to be in a given interval is approximately equal to the probability for X to be in the same interval, as n grows.
Example 3 *Let $X_1, \ldots, X_n$ be iid continuous random variables. Then trivially $X_n \rightsquigarrow X_1$. But note $P(X_1 = X_n) = 0$.*

★ *It is their distributions that are the same, not their values which are controlled by different randomness.*

Example 4 *Let $X_2, \ldots, X_n$ be identical (but not independent) copies of $X_1 \sim N(0,1)$. Then $X_n \rightsquigarrow Y = -X_1$.*

Example 5 *Let $X_n \sim \text{uniform}[0, 1/n]$. Then $X_n \rightsquigarrow \delta_0$. This is often written as $X_n \rightsquigarrow 0$.*

★ *Interestingly, note $F(0) = 1$ but $F_{X_n}(0) = 0$ for all n, so $\lim_{n\to\infty} F_{X_n}(0) \neq F(0)$. This does not contradict the definition of convergence in distribution, because $t = 0$ is not a point at which F is continuous.*

Example 6 *Let X_n has the PDF $f_n(x) = (1 - \cos(2\pi n x))$, $x \in (0,1)$. Then $X_n \rightsquigarrow \text{uniform}(0,1)$.*

★ *Note the PDFs f_n's do not converge at all, but the CDFs do.*

Theorem 1 *(The Central Limit Theorem). Let $X_1, \ldots, X_n$ be iid with finite mean μ and finite variance $\sigma^2 > 0$. Let $\bar{X}_n = \frac{1}{n}\sum_i^n X_i$. Then*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1). \tag{24}$$

Example 7 *Let $X_i \sim \text{uniform}(-1,1)$ for $i = 1, \ldots$ Note $\mu = 0, \sigma^2 = 1/3$. Then $Z_n = \sqrt{\frac{3}{n}}\sum_i X_i \rightsquigarrow N(0,1)$.*

Example 8 *Let $X_i \sim \text{beta}(\frac{1}{2}, \frac{1}{2})$. Note $\mu = \frac{1}{2}, \sigma^2 = \frac{1}{8}$. Then $Z_n = \sqrt{8n}\left(\frac{1}{n}\sum_i X_i - 1/2\right) \rightsquigarrow N(0,1)$.*

★ *The Central Limit Theorem is the most widely used application of convergence in distribution. Demo: CLT_uniform.m and CLT_beta.m*

★ *Does the CLT apply to the Cauchy distribution?*

X_n **converges to X in probability**, written $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$

$$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0. \tag{25}$$

★ *Convergence in probability is central to machine learning because a very important concept, consistency, is defined using it.*

Convergence in probability implies convergence in distribution. The reverse is not true in general. However, convergence in distribution to a point mass distribution implies convergence in probability.

★ *Example 3, Example 4, and Example 6 do not converge in probability. Example 5 does.*

★ *The definition might be easier to understand if we rewrite it as*

$$\lim_{n\to\infty} P\left(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0.$$

That is, the fraction of outcomes ω on which X_n and X disagree must shrink to zero. When $X_n(\omega)$ and $X(\omega)$ do disagree, they can differ by a lot in value. More importantly, note that $X_n(\omega)$ does not need to converge to $X(\omega)$ pointwise for any ω. This will be the distinguishing property between converge in probability and converge almost surely (to be defined shortly).

Example 9 *$X_n \xrightarrow{P} X$ in Example 2.*

X_n **converges almost surely to** X, written $X_n \xrightarrow{as} X$, if

$$P\left(\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1 \tag{26}$$

Example 10 *Example 2 does not converge almost surely to X there. This is because $\lim_{n\to\infty} X_n(\omega)$ does not exist for any ω. Pointwise convergence is central to $\xrightarrow{as}$.*

$\xrightarrow{as}$ implies $\xrightarrow{P}$, which in turn implies $\rightsquigarrow$.

Example 11 *$X_n \xrightarrow{as} 0$ (and hence $X_n \xrightarrow{P} 0$ and $X_n \rightsquigarrow 0$) does not imply convergence in expectation $\mathbb{E}(X_n) \to 0$. To see this, let*

$$P(X_n) = \begin{cases} 1/n, & \text{if } X_n = n^2 \\ 1 - 1/n, & \text{if } X_n = 0 \end{cases} \tag{27}$$

Then $X_n \xrightarrow{as} 0$. However, $\mathbb{E}(X_n) = \frac{1}{n}n^2 = n$ does not converge.

X_n **converges in rth mean** where $r \geq 1$, written as $X_n \xrightarrow{L^r} X$, if

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^r) = 0. \tag{28}$$

$\xrightarrow{L^r}$ implies $\xrightarrow{P}$.

★ *$\xrightarrow{L^r}$ implies $\xrightarrow{L^s}$, if $r > s \geq 1$. There is no general order between $\xrightarrow{L^r}$ and $\xrightarrow{as}$.*

Theorem 2 *(The Weak Law of Large Numbers). If $X_1, \ldots, X_n$ are iid, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu(X_1)$.*

Theorem 3 *(The Strong Law of Large Numbers). If $X_1, \ldots, X_n$ are iid, and $\mathrm{E}(|X_1|) < \infty$, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{as} \mu(X_1)$.*