STAT–S 782 | 0 — Scribe template | 7 August 2017

Lecturer: Prof. McDonald | Scribe: Prof. McDonald

Introduction

This document provides some information for creating the scribe files for STAT–S 682. Before you begin, please read through the macros and commands above. You should change all the `to be entered` commands as appropriate. Then, simply begin your notes in this document.

The first thing to note is that there are a number of macros which I have already defined. You should use these as much as possible. These include macros for $\mathbb{R}$, $\mathbb{E}[]$, $\mathrm{Var}[]$, and many others. Whenever you find yourself needing these symbols, please use the macros.

Second, please place your newly defined macros in the appropriate location in the header.

Theorems, definitions, conjectures, corollaries, etc have already been created using $\mathcal{AMS}$-LaTeX.

All displayed equations should be numbered for further reference.

A few notes about references: try to use the `\autoref` rather that `\ref`. That way you can type `\autoref{thm:whatever}` to get Theorem 0.2 rather than writing `Theorem \ref{thm:whatever}` to get Theorem 0.2.

Finally, please use BibTeXto create references. Simply add new references to the `.bib` file in the repo. Also, I recommend using the `natbib` package which gives, for instance, (Vapnik, 1998) or Vapnik (1998) rather than Vapnik (1998)

The rest of this document is an example from an old course. Note especially Algorithm 1.

0.1 Random variables

A *random variable* is a map X from a probability space Ω to $\mathbb{R}$. We write

$$P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}) \tag{0.1}$$

and we write $X \sim P$ to mean that X has distribution P.

The *cumulative distribution function* (cdf) of X is

$$F_X(x) = F(x) = P(X \leq x). \tag{0.2}$$

If X is discrete, its probability mass function (pmf) is

$$p_X(x) = p(x) = P(X = x). \tag{0.3}$$

If X is continuous, then its probability density function function (pdf) satisfies

$$P(X \in A) = \int_A p_X(x)dx = \int_A p(x)dx \tag{0.4}$$

and $p_X(x) = p(x) = F'(x)$. The following are all equivalent:

$$X \sim P, \qquad X \sim F, \qquad X \sim p, \qquad \mathcal{L}(X) = P. \tag{0.5}$$

Suppose that $X \sim P$ and $Y \sim Q$. We say that X and Y have the same distribution if

$$P(X \in A) = Q(Y \in A) \tag{0.6}$$

for all A. In other words, $P = Q$. In that case we say that X and Y are equal in distribution and we write $X \overset{d}{=} Y$ or $\mathcal{L}(X) = \mathcal{L}(Y)$. It can be shown that $X \overset{d}{=} Y$ if and only if $F_X(t) = F_Y(t)$ for all t.

0.2 Expected values

The *mean* or expected value of $g(X)$ is

$$\mathbb{E}[g(X)] = \int g(x) dF_X(x) = \int g(x) dP(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx & \text{if } X \text{ is continuous} \\ \sum_j g(x_j)p(x_j) & \text{if } X \text{ is discrete.} \end{cases} \tag{0.7}$$

Recall the following useful properties of expectation:

1. $\mathbb{E}\left[\sum_{j=1}^k c_j g_j(X)\right] = \sum_{j=1}^k c_j \mathbb{E}[g_j(X)]$.
2. If $X_1, \ldots, X_n$ are independent then $\mathbb{E}[\prod_{i=1}^n X_i] = \prod_i \mathbb{E}[X_i]$.
3. We often write $\mu = \mathbb{E}[X]$.
4. $\sigma^2 = \mathrm{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right]$is the Variance.
5. $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mu^2$.
6. If $X_1, \ldots, X_n$ are independent then $\mathrm{Var}[\sum_{i=1}^n a_i X_i] = \sum_i a_i^2 \mathrm{Var}[X_i]$
7. The covariance is $\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y$ and the correlation is $\rho(X, Y) = \mathrm{Cov}[X, Y] / \sigma_X \sigma_Y$. Recall that $1 \leq \rho(X, Y) \leq 1$.

The conditional expectation of Y given X is the random variable $\mathbb{E}[Y|X]$ whose valeu, when $X = x$ is $\mathbb{E}[Y|X = x] = \int y p(y|x) dy$ where $p(y|x) = p(x, y)/p(x)$. The *Law of Total Expectation* or *Law of Iterated Expectation:*

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \int \mathbb{E}[Y|X = x]\, p_x(x) dx. \tag{0.8}$$

0.3 Important distributions

Normal $X \sim N(\mu, \sigma^2)$ if

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \tag{0.9}$$

Algorithm 1 Approximate Kalman filter

```
**Input:** Initial state $w_0$, state variance $\mathbf{P}_0$, and sketching matrices $\Pi \in \mathbb{R}^{n\times m}$ and $\mathbf{R} \in \mathbb{R}^{p\times q}$.
**Compress:** $\tilde{\mathbf{H}} \to \Pi^\top \mathbf{H}\Pi$, $\tilde{\mathbf{G}} \to \mathbf{R}^\top \mathbf{G}\mathbf{R}$, $\tilde{\mathcal{X}} \to \Pi\mathcal{X}\mathbf{R}$, $\tilde{\mathbf{A}} = \mathbf{R}^\top\mathbf{A}\mathbf{R}$, $\tilde{\mathbf{P}}_0 = \mathbf{R}^\top\mathbf{P}_0\mathbf{R}$, $z_0 = \mathbf{R}w_0$.
**for** $i = 1$ **to** $T$ **do**
    Compress the data $\tilde{Y}_i = \Pi Y_i$
    Run the standard Kalman filter to produce: $\tilde{z}_i \leftarrow \tilde{\mathbf{A}}\tilde{z}_{i-1} + \tilde{\mathbf{K}}_i(\tilde{Y}_i - \tilde{\mathcal{X}}\tilde{\mathbf{A}}\tilde{z}_{i-1})$
    Update $\tilde{w}_i \leftarrow \mathbf{R}\tilde{z}_i$ and increment $\mathbf{P}_i = \mathbf{R}\tilde{\mathbf{P}}_{i-1}\mathbf{R}^\top$
**end for**
Return $w_i$ and $\mathbf{P}_i$, $i = 1, \ldots, T$.
```

Multivariate normal For $X \in \mathbb{R}^d$, $X \sim N_d(\mu, \Sigma)$ if

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\} \tag{0.10}$$

Bernoulli $X \sim Bernoulli(\theta)$ if $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$. Thus the pdf is

$$p(x; \theta) = \theta^x (1-\theta)^{1-x} I_{\{0,1\}}(x). \tag{0.11}$$

Binomial $X \sim Binomial(\theta)$ if

$$p(x; \theta) = P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} I_{\{0,\ldots,n\}}(x). \tag{0.12}$$

Uniform $X \sim U(a, b)$ if $p(x) = I_{[a,b]}(x)/(b-a)$.

Much of statistical machine learning is concerned with showing asymptotic, or better yet, finite sample properties of computational methods. We want to know whether our methods will perform well, and if so how well will they perform relative to other methods. For that, we need to discuss some useful results from probability theory. First, we'll review the convergence of random variables.

0.4 Convergence

Let $X_1, X_2, \ldots$ be a sequence of random variables, and let X be another random variable with distribution P. Let F_n be the cdf of X_n and let F be the cdf of X.

1. X_n converges *almost surely* to X, $X_n \xrightarrow{as} X$, if for every $\epsilon > 0$,

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1. \tag{0.13}$$[1]

2. X_n converges *in probability* to X, $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1. \tag{0.14}$$

3. X_n converges *in L_p* to X, $X_n \xrightarrow{L_p} X$, if

$$\lim_{n\to\infty} \int |X_n - X|^p dP = \lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0. \tag{0.15}$$

[1] The absolute value here can be replaced by any appropriate distance.

4. X_n converges *in distribution* to X, $X_n \rightsquigarrow X$, if
$$lim_{n\to\infty} F_n(t) = F(t) \tag{0.16}$$
for all t for which F is continuous.

Example 0.1. *This example shows that convergence in probability does not imply almost sure convergence. Let $S = [0,1]$. Let P be uniform on $[0,1]$. We draw SP . Let $X(s) = s$ and let*

$$\begin{array}{lll}
X_1 = s + I_{[0,1]}(s) & X_2 = s + I_{[0,1/2]}(s) & X_3 = s + I_{[1/2,1]}(s) \\
X_4 = s + I_{[0,1/3]}(s) & X_5 = s + I_{[1/3,2/3]}(s) & X_6 = s + I_{[2/3,1]}(s)
\end{array}$$

etc. Then $X_n \xrightarrow{P}$ since $P(|X_n - X| > \epsilon)$ is equal to the probability of an interval of s values whose length is going to zero. However, for every s, $X_n(s)$ alternates between the values s and $s+1$ infinitely often, so this convergence does not occur almost surely.

Theorem 0.2. *The following relationships hold:*

(a) $X_n \xrightarrow{L_p} X$ implies that $X_n \xrightarrow{P} X$.

(b) $X_n \xrightarrow{P} X$ implies that $X_n \rightsquigarrow X$.

(c) If $X_n \rightsquigarrow X$ and if $P(X = c) = 1$ for some real number c, then $X_n \xrightarrow{P} X$.

(d) $X_n \xrightarrow{as} X$ implies that $X_n \xrightarrow{P} X$.

Theorem 0.3. *Let X_n, X, Y_n, Y be random variables. Let g be a continuous function. Let c be a constant.*

(a) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.

(b) If $X_n \xrightarrow{L_p} X$ and $Y_n \xrightarrow{L_p} Y$, then $X_n + Y_n \xrightarrow{L_p} X + Y$.

(c) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n + Y_n \rightsquigarrow X + c$.

(d) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_nY_n \xrightarrow{P} XY$.

(e) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_nY_n \rightsquigarrow cX$.

(f) If $X_n \xrightarrow{\textit{converges somehow}} X$, then $g(X_n) \xrightarrow{\textit{converges the same}} g(X)$.

- *Parts (c) and (e) are known as* Slutsky's theorem*.*
- *Part (f) is known as* The continuous mapping theorem*.*

0.5 LLNs and CLT

Theorem 0.4 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be independent random variables, each with finite mean and variance. Define $S_n = X_1 + \cdots + X_n$. Then $\frac{1}{n}(S_n - \mathbb{E}[S_n]) \xrightarrow{P} 0$.*

Theorem 0.5 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be iid random variables with common mean m. Define $S_n = X_1 + \cdots + X_n$. Then $S_n/n \xrightarrow{as} m$.*

The laws of large numbers tell us that the probability mass of an average of random variables "piles up" near its expectation. In just a minute, we will see even more: how fast this piling occurs. But first we should talk about the distribution of the average.

Theorem 0.6 (Central limit theorem)**.** *Let* $X_1, X_2, \ldots$ *be iid with mean* μ *and variance* $\sigma^2 < \infty$*. Let* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$*. Then,*

$$Z_n := \frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}\left[\overline{X}_n\right]}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z, \tag{0.17}$$

where $Z \sim N(0, 1)$*.*

References

Vapnik, V. (1998), *Statistical learning theory*, John Wiley & Sons, Inc., New York.