ECE313: Probability

Sparksnotes of Prof. Hajek's sparksnotes.

Pradyun Narkadamilli

Contents

1 Midterm 1

1.1 Random Fundamentals

- An experiment can be modeled by a probability space, which has 3 terms
  - Ω - sample space. set of possible outcomes.
  - $\mathbb{F}$ - set of subsets of Ω called *events*
    - $\Omega \in \mathbb{F}$, $\in \mathbb{F}$
    - If A is an event, then $\overline{A}$ is *also* an event
    - $A, B \in \mathbb{F} \rightarrow A \cup B \in \mathbb{F}$
    - If the sample lands within the event subset, the event is considered true
  - P assigns a probability $P(A)$ to each event $A \in \mathbb{F}$
    - $P(\Omega) = 1$
    - $P(A \cup B) = P(A) + P(B) - P(AB)$
      - For mutually exclusive events A, B, $P(A \cup B) = P(A) + P(B)$
- $\binom{n}{k} = \frac{n!}{(n-k)!k!} = \binom{n}{n-k}$

1.2 Discrete Random Variables

- A random variable is considered *discrete* if it can take on a countably infinite number of values
- The expectation $E[X] = \sum_i u_i p_X(u_i)$
- The variance $\sigma^2 = (Var(X)) = E[X^2] - (E[X])^2 = E[(X - \mu_X)^2]$
  - $(Var(aX + b)) = a^2(Var(X))$
- *Conditional Probability* - $P(B|A) = \frac{P(AB)}{P(A)}$ assuming that $P(A) > 0$
- Events are considered independent if $P(AB) = P(A)P(B)$
  - Random Variables are considered indepndent if all of their events are independent w.r.t events of the other RV

1.3 Types of Discrete RVs

- *Bernoulli Distribution* - takes parameter p, returns 1 with probability p and 0 with probability $1 - p$
  - $E[X] = p$
  - $\mathrm{Var}(X) = p(1 - p)$
- *Binomial Distribution* - takes parameters n, p. Denotes the sum of n independent bernoulli trials all with parameter p
  - $p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}$
  - $E[X] = np$
  - $\mathrm{Var}(X) = np(1 - p)$
- *Geometric Distribution* - if you repeatedly run Bernoulli trials with parameter p, how many trials before your first success

- $p_L(k) = (1-p)^{k-1}p$, $k \geq 1$
- $P\{L > k\} = (1-p)^k$, for $k \geq 0$
- $E[L] = \frac{1}{p}$
- $\text{Var}(L) = \frac{1-p}{p^2}$
- *Memoryless Property* - If we seek the probability that we succeed in $k + n$ trials, but n trials have already passed, it is the same as saying we want to know the probability of succeeding k trials from the beginning.

- *Bernoulli Process* - infinite sequence of Bernoulli RVs with same p parameter
  - (X_i) is the vector of samples, (L_i) is the vector of how many trials it took before the i-th success, (C_i) is the number of successes by the i-th trial, and S_i is the number of trials required for the i-th success

- *Negative Binomial Distribution* - estimates the number of trials n required to find r successes with Bernoulli parameter p
  - $p(n) = \binom{n-1}{r-1}p^r(1-p)^{n-r}, n \geq r$
  - $E[S_r] = \frac{r}{p}$
    * $\text{Var}(S_r) = \frac{r(1-p)}{p^2}$

- *Poisson Distribution* - limit of binomial distribution. Becomes a good approximation when n is very large and p is very small
  - Poisson parameter $\lambda = np$, where n, p are classic Binomial parameters
  - $p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$
  - $E[N] = \lambda$, $\text{Var}(N) = \lambda$

1.4 More Random Stuff

- *Maximum Likelihood Parameter Estimation* - we select a parameter such that the probability of the outcome is maximized.

- *Markov Inequality* - For a nonnegative random variable Y, $P\{Y \geq c\} \leq \frac{E[Y]}{c}$

- *Chebychev Inequality* - $P\{p \in (\widehat{p} - \frac{a}{2\sqrt{n}}, \widehat{p} + \frac{a}{2\sqrt{n}})\} \geq 1 - \frac{1}{a^2}$
  - With the above inequality, we can derive expressions on different confidence intervals using confidence $= 1 - \frac{1}{a^2}$

- *Bayes' Formula* - $P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)}$
  - Each event E used as a condition for Baye's formula helps to form a partition on the conditioned event A - thus, when looking at the law of total probability, we must assert that all the events considered in the partition must be mutually exclusive

2 Midterm 2

Random stuff is discussed here, when it can't be placed within another subsection.

- **Union Bound**: $P(A \cup B) \leq P(A) + P(B)$
  - Can enumerate all potential failure scenarios in a network and add their probabilities to find union bound

2.1 Hypothesis Testing

- *Maximum Likelihood Rule* - for a given event, which hypothesis had the higher probability?
- *Maximum a Posteriori (MAP)* - one hypothesis is known to be more likely than the other

Both types take a general form as follows. We assume that π_0 and π_1 are the probabilities of either hypothesis occuring. The ML rule can be considered a special case of MP, where $\pi_0 = \pi_1 = \frac{1}{2}$.

$$\Lambda(k) = \frac{p_1(k)}{p_0(k)}$$

$$H = \begin{cases} \text{Declare } H_1, & \Lambda(k) > \frac{\pi_0}{\pi_1} \\ \text{Declare } H_0, & else \end{cases}$$

Given a decision rule, we should also consider how often it is wrong.

$$p_{falsealarm} = P(H_1 \text{ declared } | H_0 \text{ is true})$$

$$p_{miss} = P(H_0 \text{ declared } | H_1 \text{ is true})$$

$$p_e = \pi_0 p_{falsealarm} + \pi_1 p_{miss}$$

2.2 Types of Continuous Random Variables

- *Cumulative Distribution Fundation* (CDF), and *Probability Density Function* (pdf)
  - CDF refers to all the probability until current value, pdf refers to point probability of current outcome - cannot be considered absolute probability
  - $P(X = a) = 0$, always
- *Uniform Distribution*: $\text{Unif}(a, b)$: any value between a and b is equally likely
  - Mean: $\frac{a+b}{2}$
  - Variance: $\frac{(b-a)^2}{12}$
  - pdf (within bounds): $f(u) = \frac{1}{b-a}$
- *Exponential Distribution*: $\text{Exp}(\lambda)$: limit of scaled geometric random variables - memoryless property still holds. λ value determines both initial value and speed of decay rate.
  - Mean: $\frac{1}{\lambda}$
  - Variance: $\frac{1}{\lambda^2}$
  - pdf (for $u \geq 0$): $f(u) = \lambda e^{-\lambda u}$
  - CDF (for $u \geq 0$): $F(u) = 1 - e^{-\lambda u}$
- *Gaussian Distribution*: $N(\mu, \sigma^2)$: The normal boi.

- Mean and Variance are given as distribution parameters
- pdf (for $u \in \mathbb{R}$): $f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(u-\mu)^2}{2\sigma^2}}$

- *Properties of the Gaussian Distribution*
  - A normalized Gaussian Distribution is defined as $N(0, 1)$
  - Such a distribution has CDF $\phi(c)$ and inverse CDF $Q(c)$
    * $\phi(c) = Q(-c) = 1 - Q(c)$
  - Any Gaussian Distribution can be standardized to this form
    * $\tilde{X} = \frac{X-\mu}{\sigma}$ - this transformation can be performed on a per-sample value as well

- *Poisson Process*: Poisson(λ) : Models the number of counts in a time interval - acts as a collection of identically and independently distributed exponential RVs. Happens to be the limit of a scaled Bernoulli process.
  - λ must be scaled per the time interval
  - The fundamental per-timestep λ rate can be used in the exponential distribution as well
  - $\mu = \sigma^2 = \lambda$
  - pmf (for $u \in \mathbb{R}$): $p(k) = \frac{e^{-\lambda}\lambda^k}{k!}$
    * k is the expected/relevant count

- *The Erlang Distribution* - Estimates the probability that the r th count of a Poisson Process happens at/after time t.
  - PDF: $f_{T_r}(t) = \frac{e^{-\lambda t}\lambda^r t^{r-1}}{(r-1)!}$
  - CDF: $P\{T_r > t\} = \sum_{k=0}^{r-1} \frac{e^{-\lambda t}(\lambda t)^k}{k!}$
  - $\mu = \frac{r}{\lambda}$, $\text{Var}(T_r) = \frac{r}{\lambda^2}$

2.3 Properties of Continuous Random Variables

We should consider that PDFs can be scaled linearly in the same way that we manipulated discrete PMFs. Consider an arbitrary transformation $Y = aX + b$ where X is a continuous random variable.

- $E[Y] = aE[X] + b$
- $\text{Var}(Y) = a^2\text{Var}(X)$
- $f_Y(u) = \frac{1}{a} f_X(\frac{u-b}{a})$

There is also the concept of *ML parameter estimation* - given that you know some information about a distribution, and that a certain sample is observed, you must optimize the remaining parameter(s) of the CRV to maximize the likelihood that the observed sample is pulled out.

- Quite literally the same as the discrete case.

There is also the matter of performing nonlinear transformations on a random variable. Suppose we have $Y = g(X)$.

- If Y is continuous, find the CDF of Y, then differentiate to find the PDF
- If Y is discrete, then we can directly find the PMF, where $p_Y(u) = P(g(X) = u)$

- $E[Y] = \int_{-\infty}^{\infty} g(u) f_X(u) du$
- ***DO NOT SIMPLY TRANSFORM THE COORDINATES*** - you need to appropriately scale the probability density.

In a similar fashion, we can perform some transformation $g(c)$ on Unif$(0, 1)$ to produce any continuous distribution. This is usually done by inverting the taking the CDF of a given distribution and inverting it - that is to say, $g(c) = F_X^{-1}(c)$, where c would traditionally be the *probability output* of the CDF, but here we use it to find the relevant input that would have produced such a probability… weird, huh?

- **Gaussian Approximation** - we can approximate the sum of n independent random variables, where each random variable is small in magnitude compared to the sum, as a Gaussian distribution.
  - We call this idea the *Central Limit Theorem*
  - This approximation is very good with binomial distributions, where np and $n(1-p)$ are moderately large
    - In such a case, the distribution is approximated with $E[X] = np$, and $\text{Var}(X) = np(1-p)$
  - *Continuity Correction* - we apply this when we attempt to use the Gaussian Approximation with an integer-valued random variable
    - $P\{X \leq k\} \approx P\{\tilde{X} \leq k + 0.5\}$, $P\{X \geq k\} \approx P\{\tilde{X} \geq k - 0.5\}$, and $P\{X = k\} \approx P\{k - 0.5 \leq \tilde{X} \leq k + 0.5\}$
- *Failure Rate Functions* - the failure rate function for distribution T refers to the concept that if the item is working at time t, we want the probability that the item fails within the next ϵ time units. T is the distribution describing the lifetime of the item.
  - Failure Rate Function for $t \geq 0$: $h(t) = \frac{f_T(t)}{1-F_T(t)}$
  - CDF given $t \geq 0$: $F(t) = 1 - e^{-\int_0^t h(s) ds}$
  - $E[T] = \int_0^{\infty} (1 - F(t)) dt$
  - $-\ln(1-\epsilon) \approx \epsilon$ for a small ϵ value - this can be used for simplification on exams

3 Final

3.1 Jointly Distributed Random Variables

- A joint CDF will essentially bisect an N-D space into 2 halves, where the region going towards $-\infty$ in multiple variables is included.

$$F_{X,Y}(u_0, v_0) = P\{X \leq u_0, Y \leq v_0\}$$

- Probability that two RVs fall within a bounding rectangle R with bottom left corner (a, c) and top right corner (b, d)

$$P\{(X, Y) \in R\} = F(b, d) - F(b, c) - F(a, d) + F(a, c)$$

- Restating the properties of a CDF, but bivariate
  - F is nondecreasing on both u and v
  - F(u,v) is right-continuous on both its inputs - that is to say, if you approach the function from the right side, the limit of a point converges to the value at said point.
  - The probability of being within a rectangular region of the 2D variable space is non-negative
  - The limit approaching $-\infty$ on either variable is 0, the limit approaching ∞ on *both* variables is 1
- For multiple discrete variables, you have a joint probability mass function (pmf) defined such that $p_{x,y}(u, v) = P\{X = u, Y = v\}$.
  - Given a joint pmf, the univariate pmf can be found by taking the union of all events of the other variables
  - Ex: for a bivariate joint PMF, $P\{X = u\} = \sum_j P(X = u, Y = v_j) = \sum_j p_{X,Y}(u, v_j)$, where j iterates over all possible outcomes of the discrete variable Y.
- We can use virtually the same logic as above with joint probability density functions
  - For the joint CDF, we perform a double integral with an upper bound on each integral's bounds
- Suppose we take a function of the variables in our joint distribution, we can find its expected value as follows

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v) f_{X,Y}(u, v) du dv$$

  - By LOTUS, we can still use linearity of expectation when working with joint distributions
- There are two requirements that must be fulfilled for a (in this case bivariate) joint pdf to be considered valid
  - For any combination (u, v), $f_{X,Y} \geq 0$ (no such thing as negative probability)
  - The integral of the joint pdf over the entirety of the N-D space is 1
  - f_X and f_Y extracted from the joint distribution's pmf are called *marginal pmfs*
  - The *conditional pdf* $f_{Y|X}(v|u_0) = \frac{f_{X,Y}(u_0,v)}{f_X(u_0)}$, and is only defined if the u_0 in question has a nonzero pdf

- We can use the conditional pdf to extract the *conditional expectation*, $E[Y|X=u] = \int_{-\infty}^{\infty} v f_{Y|X}(v|u)dv$
    - When computing this, remember to *divide by the value of the marginal pdf*
- When dealing with joint distributions, specifically looking at a bivariate distribution, we can determine independence using the following

$$F_{X,Y}(u_0, v_0) = F_X(u_0)F_Y(v_0) \text{ or } P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$$

  - Equivalently, we note that $f_{X,Y}(u,v) = f_X(u)f_Y(v)$, or $p_{X,Y}(u,v) = p_X(u)p_Y(v)$ for discrete variables
  - We can also determine independence using conditional probabilities - the independence of variables in a joint distribution is only held if $f_X(u) > 0$, then $f_{Y|X}(v|u) = \frac{f_{X,Y}(u,v)}{f_X(u)}$
    - Conditionality has no effect on the probability
- *Swap Property* - suppose we take any two points (a,b) and (c,d) - if the rectangle formed with those points at the farthest apart corners has all 4 vertices in the set $S \in \mathbb{R}^2$, it has the swap property
  - Required in any product set (the joint distribution formed by the permutations of two different random variables)
  - For independent continuous variables X, Y, the support of $f_{X,Y}(u,v)$ is a product set of those two variables
  - For (X,Y) uniformly distributed over set S in the plane, X, Y are independent iff S is a product set.
- Realistically, you should try to reason the possibilities of permutations that yield the desired value for a function of the joint variables, then do the funny integrals
  - By coincidence, to find the probability that the sum of two independent variables takes a certain output value, you can convolve the two variables (discrete *or* continuous)
- The same "find distribution of function of a single RV" method can be used here, albeit with multivariate adaptation
  - Identify support of all inputs, identify the relevant support of the output
  - Find the CDF of the output for a given constant c
  - Derive the CDF to find the PDF of the function
- Suppose that we have W and Z, which are linear functions of X and Y (form $aX + bY$ or $cX + dY$). We can denote this as a matrix.

$$\begin{pmatrix} W \\ Z \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

If we refer to the 2x2 matrix as A, then we can assert that $f_{W,Z}(\alpha, \beta) = \frac{1}{|\det A|} f_{X,Y}(A^{-1}\binom{\alpha}{\beta})$. This basically allows us to transform one bivariate joint distribution into *another* bivariate joint distribution.

3.2 Correlation and Covariance

- Suppose we have random variables X and Y with finite variances
  - *Correlation*: $E[XY]$
  - *Covariance*: $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
    - If two random variables are *uncorrelated*, then their covariance is 0 - two independent variables can be considered uncorrelated. However *uncorrelated* variables are not necessarily independent.
  - *Correlation Coefficient*: $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_Y \sigma_X}$
    - Standardized versions of X and Y will have a covariance of $\rho_{X,Y}$
  - Note that the covariance of a variable with regards to itself is just its own variance
  - For independent variables, $E[XY] = E[X]E[Y]$, since each variable has no bearing on the other

For multiple random variables (three or more), they are considered uncorrelated if they are all *pairwise* uncorrelated.

$$\text{Cov}(X + Y, U + V) = \text{Cov}(X, U) + \text{Cov}(X, V) + \text{Cov}(Y, U) + \text{Cov}(Y, V)$$
$$\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$$

For the special case where X and Y are uncorrelated (Covariance is 0), we can do a funny ting.

$$\text{Var}(X+Y) = \text{Cov}(X+Y, X+Y) = \text{Cov}(X, X)+\text{Cov}(Y, Y)+2\text{Cov}(X, Y) = \text{Var}(X)+\text{Var}(Y)$$

For any two random variables X and Y, the following holds.

$$\|E[XY]\| \leq \sqrt{E[X^2]E[Y^2]}\text{and}\|\text{Cov}(X, Y)\| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}$$

Nice formula to have on hand about sum of variables.

$$\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i) + \sum_j \sum_{k,k\neq j} \text{Cov}(X_j, X_k)$$

3.3 Estimation of Correlated Random Variables

- *Mean Squared Error (MSE)*: $E[(Y - \widehat{Y})^2] = \int_{-\infty}^{\infty} (v - \widehat{v})^2 f_Y(v) dv$, where $\widehat{Y}$ is the estimator used for Y
  - This is a symbolic integral, and is not mathematically substitutable. One should logically add additional integral dimensions if the estimator varies between different conditional values.
- Estimator is *unbiased* if its mean is equal to parameter being estimated
- *Constant Estimator* -we always predict that Y will be σ, independent of the input value
  - MSE: $\text{Var}(Y) + (E[Y] - \sigma)^2$
  - Best constant estimator is δ
- *Unconstrained Estimator* - we generate an arbitrary function $g(u)$ such that for an observation $X = u$, we minimize error.

- Best Unconstrained Estimator: $g^*(u) = E[Y|X = u] = \int_{-\infty}^{\infty} v f_{Y|X}(v|u)dv$
- MSE for Best Estimator: $E[Y^2] - E[g^*(u)^2]$
  - Can't just square avg of $g^*(u)$ dipshit
- Generic MSE: $E[(Y - E[Y|X])^2]$

- *Linear Estimator* - better than a constant, but can be easier to compute than an unconstrained estimator
  - Optimal Linear Estimator: $L(u) = \widehat{E}[Y|X = u] = \mu_Y + \frac{\text{Cov}(Y,X)}{\text{Var}(X)}(u - \mu_X) = \mu_Y + \sigma_Y \rho_{X,Y}(\frac{X-\mu_X}{\sigma_X})$
  - Generic Linear Estimator MSE: $\text{Var}(Y) - 2a\text{Cov}(Y, X) + a^2\text{Var}(X)$
  - MSE for optimal estimator: $\sigma_Y^2(1 - \rho_{X,Y}^2) = E[Y^2] - E[\widehat{E}[Y|X]^2]$

3.4 Law of Large Numbers and CLT

- General idea is that as you increase the number of samples taken from a distribution, if you standardize that distribution based on its mean and variance, then it will appear to be a normal distribution.

- *Law of Large Numbers* - We can formulate this based on the Chebychev inequality. The expression implies that as we increase our number of observations, we get results more and more likely to be close to the mean. Suppose that we have a variable $S_n = X_1 + X_2 + X_3 + ... + X_i$ where all the X variables have mean μ, and a variance $\leq C$
  - $P\{|\frac{S_n}{n} - \mu| \geq \delta\} \leq \frac{\text{Var}(S_n/n)}{\delta^2}$, which approaches 0 as $n \to \infty$
  - $E[\frac{S_n}{n}] = \mu$
  - $\text{Var}(\frac{S_n}{n}) = \frac{C}{n}$ if the X variables are uncorrelated

- *Central Limit Theorem* - using the same variable definitions as in the previous LLN, we can declare the following
  - $\lim_{n\to\infty} P\{\frac{S_n - \mu_{S_n}}{\sqrt{\text{Var}(S_n)}} \leq c\} = \phi(c)$
  - May need to add continuity correction to the way we evalute this expression for finite n

3.5 Joint Gaussian Distribution

- Random variables X and Y are *jointly gaussian* if every $aX + bY$ is a gaussian random variable
  - Constants are considered Gaussian
  - The degenerate cases where $X = aY + b$ and $Y = aX + b$, where the indep. var is a Gaussian RV still count as being jointly Gaussian
  - For non-degenerate cases, jointly Gaussian RVs have the following PDF, assuming that variances are nonzero and the correlation coefficient is bounded to ± 1:

$$f_{X,Y}(u, v) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho^2}} \exp(-\frac{(\frac{u-\mu_X}{\sigma_X})^2 + (\frac{v-\mu_Y}{\sigma_Y})^2 - 2\rho(\frac{u-\mu_X}{\sigma_X})(\frac{v-\mu_Y}{\sigma_Y})}{2(1 - \rho^2)})$$

- A pdf is considered bivariate normal if it is of the form $f_{X,Y}(u, v) = C\exp(-P(u, v))$, and $P(u, v) = au^2 + buv + cv^2 + du + ev + f$ (second order bivariate polynomial)

  - C is selected such that the PDF integrates over the 2D plane to 1
  - Both variables in the bivariate normal distribution are considered normal on their own ($N(\mu_X, \sigma_X^2)$)
  - Any combination $aX + bY$ is a gaussian random variable
  - X and Y are independent iff $\rho = 0 \rightarrow \text{Cov}(X, Y) = 0$
  - The best estimator for Y (the unconstrained estimator) is linear
  - Conditional distribrution of Y given $X = u$ is $N(\widehat{E}[Y|X = u], \sigma_e^2)$, σ_e^2 is MSE of $\widehat{E}[Y|X]$
    * The center of this conditional probability distribution *is* the linear estimator.

- Joint Gaussian Distributions can be scaled to any number of dimensions - suppose we have N-D vectors - random vector$X$, μ for mean vector, and u for argument. X has n-dimensional Gaussian density
  - For the covariance matrix, the (i, j) entry is given by $\text{Cov}(X_i, X_j)$, with diagonal given by variance of the i-th variable
  - Assuming the conditions above, for the NxN covariance matrix Σ, PDF is given by:

$$f_X(u) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma)}} \exp(-\frac{(u-\mu)^T\Sigma^{-1}(u-\mu)}{2})$$

- If you've forgotten how to take the determinant:

$$\det(\begin{pmatrix} a & b \\ c & d \end{pmatrix}) = ad - bc$$

This scales differently at higher dimensions. We iterate along a column or row, and find the determinant for the matrix excluding the column/row. As we iterate, we invert the additive sign every iteration. An example for 3x3 matrix is shown below.

$$\det(\begin{pmatrix} a & b & c \\ e & f & g \\ h & i & j \end{pmatrix}) = a(fj - gi) - b(ej - gh) + c(ei - fh)$$