MSML610: Advanced Machine Learning

Probability Distributions

Instructor: GP Saggese, PhD - gsaggese@umd.edu

References:

Interesting RVs

- **Interesting RVs**
  - Bernoulli
  - Binomial
  - Gaussian
  - Log-Normal
  - Poisson
  - Chi-square
  - Student's t-distribution
- Probability inequalities

Bernoulli

- Interesting RVs
  - **Bernoulli**
  - Binomial
  - Gaussian
  - Log-Normal
  - Poisson
  - Chi-square
  - Student's t-distribution
- Probability inequalities

Bernoulli distribution: definition

- The Bernoulli distribution

  $$X \sim \text{Bernoulli}(p)$$

  represents flipping a coin that has probability $0 \leq p \leq 1$ of coming up heads:

  $$\Pr(X = 1) = p$$

  $$\Pr(X = 0) = 1 - p$$

Bernoulli distribution: PDF

- The PDF can be written in terms of a function:

$$f_X(x) = \Pr(X = x) = p^x(1-p)^{1-x} \text{ for } x \in \{0, 1\}$$

Bernoulli distribution: mean and variance

- Given $X \sim \text{Bernoulli}(p)$
- Mean: $\mathbb{E}[X] = p$
- Variance: $\mathbb{V}[X] = p(1 - p)$

Binomial

- Interesting RVs
  - Bernoulli
  - **Binomial**
  - Gaussian
  - Log-Normal
  - Poisson
  - Chi-square
  - Student's t-distribution
- Probability inequalities

Binomial distribution: definition

- The Binomial distribution

$$X \sim \text{Binomial}(n, p)$$

represents the number of heads when tossing n times a coin with probability of heads p:

$$\Pr(X = x) = \Pr(\text{getting } x \text{ heads}), 0 \leq x \leq n$$

Binomial distribution: PDF

- The PDF is

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x \in \{0, ..., n\}$$

- In fact one can evaluate this probability as the sum of the probability of all possible events, since they are all mutually exclusive:
  - Each event has probability $p^x(1-p)^{n-x}$ (because it is a Bernoulli trial)
  - There are $\binom{n}{x}$ possible ways of choosing a set of x objects (in this case heads) out of n identical objects (the binomial coefficient)

Relationship between Binomial and Bernoulli variables

- A Binomial variable $Y \sim \text{Binomial}(n, p)$ can be written as sum of n IID Bernoulli variables $X_i \sim \text{Bernoulli}(p)$:

$$Y = \sum_{i=1}^{n} X_i$$

Mean and variance of Binomial distribution

- By using the relationship between Binomial and Bernoulli variables

$$X \sim \text{Binomial}(n, p) = \sum_{i=1}^{n} X_i$$

- Mean: $\mathbb{E}[X] = np$
- Variance: $\mathbb{V}[X] = np(1-p)$

Example of binomial distribution (7 girls)

- A friend has 8 children, 7 of which are girls
- What's the probability of this event, if each birth is independent and each gender has 50%-50%?
- ***Solution***
- We can look at that as 8 configurations (boy in *i*-st position) out of 2^8, which is $\frac{1}{2^5}$
- Using binomial distribution, head is girl then

$$\Pr = \binom{8}{7}(1/2)^7(1/2) = \frac{1}{2^5}$$

- What's the probability of having at least 7 girls?
- It is the probability of having 7 or 8 girls, i.e.

$$\Pr = \binom{8}{7}(1/2)^7(1/2) + \binom{8}{8}(1/2)^8 = (\binom{8}{7} + \binom{8}{8})(1/2)^8 = \binom{9}{8}(1/2)^8 = \binom{9}{}$$

Gaussian

- Interesting RVs
  - Bernoulli
  - Binomial
  - **Gaussian**
  - Log-Normal
  - Poisson
  - Chi-square
  - Student's t-distribution
- Probability inequalities

Gaussian distribution

- Aka Normal
- A variable is Gaussian

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

has a PDF in the form:

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

- MEM: note that we use the variance σ^2 to avoid square roots coming from std dev

Gaussian distribution: parameters

- Given $X \sim N(\mu, \sigma^2)$
- Mean: $\mathbb{E}[X] = \mu$
- Variance $\mathbb{V}[X] = \sigma^2$

A standard Gaussian

- $=$ a Gaussian $\sim N(0,1)$, indicated with Z

Area under the center of a Gaussian curve

- Consider a Gaussian $N(\mu, \sigma^2)$
- The area under the center of the curve is:
  - $\Pr(|X - \mu| \leq \sigma) = 68\%$
  - $\Pr(|X - \mu| \leq 2\sigma) = 95\%$
  - $\Pr(|X - \mu| \leq 3\sigma) = 99\%$
- MEM: 68-95-99

Area under 2 tails of a Gaussian curve

- Consider a Gaussian $N(\mu, \sigma^2)$
- The area under each tail can be derived from the 68-95-99 numbers by considering the difference with 100%
- The area under the two tails is:
  - $\Pr(|X - \mu| \geq \sigma) = 100 - 68\% = 32\%$
  - $\Pr(|X - \mu| \geq 2\sigma) = 100 - 95\% = 5\%$
  - $\Pr(|X - \mu| \geq 3\sigma) = 100 - 99\% = 1\%$
- MEM: 68-95-99

Area under 1 tail of a Gaussian

- The area under each tail can be derived from the 68-95-99 numbers by dividing by 2 the difference with 100%
  - $\Pr(X > \mu + \sigma) = 16\%$
  - $\Pr(X > \mu + 2\sigma) = 2.5\%$
  - $\Pr(X > \mu + 3\sigma) = 0.5\%$

1-side Gaussian quantiles for $1, 2, 3\sigma$

- From the previous numbers we can get some approximated quantiles:
  - 1σ from μ corresponds to 68% + 16% = 84%
  - 2σ from μ corresponds to 95% + 2.5% = 97.5%
  - 3σ from μ corresponds to 99% + 0.5% = 99.5%
- Remember that α-quantile is defined as the value x_α such that $\Pr(X < x_\alpha) = \alpha$ (i.e., the area under the portion of the curve is α)
- These by definition are the 84%, 97.5% and 99.5% quantiles
  - $x_{0.84} = \mu + 1\sigma$
  - $x_{0.975} = \mu + 2\sigma$
  - $x_{0.995} = \mu + 3\sigma$

1-side Gaussian quantiles for $95\%, 97.5\%, 99\%$

- $x_{0.95} = \mu + 1.645\sigma$
- $x_{0.975} = \mu + 1.96\sigma$
- $x_{0.99} = \mu + 2.33\sigma$
- MEM: 1.645, 1.96, 2.33 for 95, 97.5, 99

1-sided and 2-sided quantiles for symmetric distributions

- For symmetric distributions the 1-sided and 2-sided quantiles are related by the relationship:

$$q_{2s}(\alpha) = q_{1s}(\frac{1+\alpha}{2})$$

1-sided and 2-sided quantiles for symmetric distributions: proof

- This proof holds for any symmetric around 0 PDF
- Consider that

$$CDF_{2s}(x) = CDF_{1s}(x) - CDF_{1s}(-x)$$

  given the definition of CDF in terms of integral of a PDF
- Given the symmetry of the PDF it holds:

$$CDF_{1s}(-x) = 1 - CDF_{1s}(x)$$

- Thus

$$CDF_{2s}(x) = 2 \cdot CDF_{1s}(x) - 1$$

- Now we need to express the 2-sided quantile in terms of 1-sided quantile where the quantile is the inverse of the corresponding CDF
- If $f(x) = g(h(x))$ with both g and h invertible, then

$$f^{-1}(x) = h^{-1}(g^{-1}(x))$$

- The inverse of $y = 2x - 1$ is $x = (y + 1)/2$ and the inverse of CDF_{1s} is q_{1s}, therefore

$$q_{2s}(x) = q_{1s}((x + 1)/2)$$

What is the 95% percentile of a Gaussian?

- Given a Gaussian $X \sim N(\mu, \sigma^2)$, what is the 95% percentile?
- ***Solution***
- The 95% percentile is by definition the value $x_{0.95}$ of the RV such that $\Pr(X \leq x_{0.95}) = 0.95$
- This value is $\mu + 1.645\sigma$

Example of computing probabilities with Gaussian

- The daily ad-clicks for a company are approximately distributed as $X \sim N(\mu = 1020, \sigma^2 = 50^2)$
- What's the probability of getting more than 1,160 clicks?
- ***Solution***
- $\Pr(X \geq 1160) = 1 - F_X(1160) = 1 - F_Z((1160 - 1020)/50) = 1 - F_Z(2.8)$
- It's not very likely since $\Pr(Z \geq 2.33) = 1 - \Pr(Z \leq 2.33) = 1 - 99\% = 1\%$

Log-Normal

- Interesting RVs
    - Bernoulli
    - Binomial
    - Gaussian
    - **Log-Normal**
    - Poisson
    - Chi-square
    - Student's t-distribution
- Probability inequalities

Log-normal distribution

- A RV has a log-normal distribution

$$X \sim LN(\mu, \sigma^2)$$

  if the log of the RV is Gaussian: $\log(X) \sim N(\mu, \sigma^2)$
- A log-normal RV assumes only positive values

Log-normal: mean and variance

- Given $X \sim LN(\mu, \sigma^2)$
- Mean: $\mathbb{E}[X] = \exp(\mu + \sigma^2)$
- Variance: $\mathbb{V}[X] = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$
- TODO: look at the PDF shape

Poisson

- Interesting RVs
    - Bernoulli
    - Binomial
    - Gaussian
    - Log-Normal
    - **Poisson**
    - Chi-square
    - Student's t-distribution
- Probability inequalities

Poisson distribution: interpretation

- A RV $X \sim \text{Poisson}(\lambda)$ models counts or arrivals per unit of time
  - It represents the probability of getting x arrivals in a unit of time
- It is a discrete and positive RV

Poisson distribution: PDF

- Its PDF is:

$$f_X(k) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

  where k is an integer ≥ 0
- MEM:
  - It is like a prob of occurrence of k events (λ^k)
  - Divided by the number of permutations ($k!$)
  - Finally $e^{-\lambda}$ is to get the area to 1
- The PDF starts at $e^{-\lambda}$, increases, and then decreases

Poisson distribution: mean and variance

- Mean: $\mathbb{E}[X] = \lambda$
- Variance: $\mathbb{V}[X] = \lambda$
- Thus to model something as a Poisson, mean and variance need to be equal
  - This assumption can be checked using the data

Poisson distribution for modeling rates

- In practice we use Poisson scaled by the monitoring time

$$X \sim \text{Poisson}(\lambda T)$$

  where T is the total monitoring time
- Applying the expectation

$$\lambda = \frac{\mathbb{E}[X]}{T}$$

  thus λ is the average count per unit time, i.e., the rate of occurrence

Example of use of Poisson distribution

- The number of people at the bus stop is Poisson with mean of 2.5 per hour
- What's the probability that 3 or fewer people take the bus in 4 hours?
- ***Solution***
- $\Pr(X \leq 3)$ with $X \sim \text{Poisson}(2.5 \times 4)$

Poisson as approximation to the Binomial

- When $n \gg 1$ and $p \ll 1$ (i.e., many attempts of a unlikely event), then

$$\text{Binomial}(n, p) \approx \text{Poisson}(np)$$

$$Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{(np)^x e^{-np}}{x!}$$

- Note that np is the average number of successes in n attempts, since

$$p = \frac{n_{\text{success}}}{n} \implies np = n_{\text{success}}$$

  so it is like considering successes as number of arrivals in a given period

- Poisson is simpler to compute since it has only one factorial, instead of two factorials like the Binomial

Chi-square

- Interesting RVs
  - Bernoulli
  - Binomial
  - Gaussian
  - Log-Normal
  - Poisson
  - **Chi-square**
  - Student's t-distribution
- Probability inequalities

Chi-square distribution

- We say that X is "chi-square with q degrees of freedom"

$$X \sim \chi^2_q$$

$\iff$ the RV

$$X = Z_1^2 + ... + Z_q^2$$

where $Z_1, ..., Z_q$ are q independent standard Gaussians $\mathrm{N}(0, 1)$
- MEM: χ^2_q is the sum of q IID squared standard normals

Chi-square distribution: PDF properties

- X chi-square is always non-negative, i.e., $X \geq 0$
- Its PDF always from 0
- For small q degrees of freedom, it has a peak $< q$ (mode) and a long tail
- For $q \to \infty$ it is asymptotically Gaussian: χ^2_q

Chi-square

- The mean is equal to the number of degree of freedom: $\mathbb{E}[X] = q$
- It can be shown by its definition in terms of Gaussians

Student's t-distribution

- Interesting RVs
  - Bernoulli
  - Binomial
  - Gaussian
  - Log-Normal
  - Poisson
  - Chi-square
  - **Student's t-distribution**
- Probability inequalities

Student's t-distribution: definition in terms of Gaussians

- We say that T is "Student t-distribution with q degrees of freedom")

  $$T \sim t_q$$

  iff

  $$T = \frac{Z}{\sqrt{\frac{Z_1^2 + \ldots + Z_q^2}{q}}}$$

  where Z and $Z_1, \ldots, Z_q$ are $q + 1$ independent standard Gaussians $\mathrm{N}(0, 1)$
- Note that at the denominator there is the square root of a chi-square
- MEM: it's the ratio between a standard normal Z and the norm of a vector of standard normals $\sqrt{Z_1^2 + \ldots + Z_q^2}$

Student's t-distribution: definition in terms of Chi-square

- Given a Student's t-distribution with q degrees of freedom $T \sim t_q$, then

$$T = \frac{Z}{\sqrt{\frac{X}{q}}}$$

where $Z \sim N(0,1), X \sim \chi^2_q$ and independent

Student's t-distribution: shape and properties

- A t-distribution is centered around zero
  - Its mean is equal to 0 (since it is symmetric with respect to 0)
- It has thicker tails than a normal distribution
- For large q the t-distribution converges to a normal distribution
- MEM: It's like a standard Gaussian with heavier tails

Probability inequalities

- Interesting RVs
- **Probability inequalities**

PAC statements

- $=$ Probably Approximately Correct statement
- In practice there is an approximation that holds with a certain probability
- Many probability inequalities are PAC statements

Markov inequality

- ***Hypothesis***
- Given X discrete or continuous RV
- X is a non-negative RV (i.e., $X \geq 0$, PDF is all after 0)
- X has finite mean: $\mathbb{E}[X] < \infty$
- ***Thesis***
- The probability that X is larger than a certain value is bounded by the mean

$$\Pr(X \geq x) \leq \frac{\mathbb{E}[X]}{x}$$

Markov inequality: geometric interpretation

- Given a RV $X \geq 0$ with a finite mean
- The "flipped CDF" $1 - F_X(x)$ is dominated by an hyperbole passing by $(y, x) = (\mathbb{E}[X], 1)$
- This is also related to the fact that a PDF needs to sum to 1 and thus needs to decrease at least like $1/n$

Proof of Markov inequality

- TODO: Add

Chebyshev inequality

- ***Hypothesis***
- Given X discrete or continuous RV
- X with finite mean μ and variance σ^2
- ***Thesis***
- The probability that X is far from the mean is bound by the variance:

$$\Pr(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Chebyshev inequality in terms of z-scores

- ***Hypothesis***
- Given X discrete or continuous RV
- X with finite mean μ and variance σ^2
- ***Thesis***
- Expressing the distance from the mean in terms of standard deviation $\varepsilon = k\sigma$:

$$\Pr(\frac{|X-\mu|}{\sigma} \geq k) \leq \frac{1}{k^2}$$

- The probability that the z-score of a RV is far away from 0 at least a certain number k is bounded by $\frac{1}{k^2}$

Proof of Chebyshev inequality

- TODO: Add

Comparing Markov and Chebyshev inequalities

- Markov assumes $X \geq 0$
- Chebyshev makes no assumptions
- Both inequalities have a similar form:

$$\Pr(X \geq x) \leq \frac{\mu}{x}$$

$$\Pr(|X - \mu| \geq x) \leq \frac{\sigma^2}{x^2}$$

Hoeffding inequality

- Given a Bernoulli RV with probability of success μ
- We want to estimate μ using N samples:

$$\nu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- Then

$$\Pr(|\nu - \mu| > \varepsilon) \leq 2e^{-2\varepsilon^2 N}$$

- Since ν is bound in $[\mu - \varepsilon, \mu + \varepsilon]$, we want a small ε with a large probability