COE241 - Estatística e Modelos Probabilísticos
COS868 - Probabilidade e Estatística para Aprendizado de Máquina
Segundo Semestre de 2025 - Professora: Rosa Maria Meri Leão

Projeto do Curso - Primeira Parte

1 Objetivo

O objetivo deste trabalho é aplicar a teoria de Estatística, Máxima Verossimilhança e Inferência Bayesiana a um conjunto de dados real. É essencial que seja realizada uma análise e interpretação dos resultados encontrados em todas as etapas.

O *dataset* a ser analisado contém medições de parâmetros de conexões da Internet. As análises incluem:

- Calcular **estatísticas descritivas** e elaborar **gráficos exploratórios**.
- Ajustar modelos paramétricos utilizando a **Máxima Verossimilhança (MLE)**.
- Realizar **inferência bayesiana** com *priors conjugadas* para obter as *posteriors* e previsões.
- Comparar as estimativas **MLE vs Bayes**.

2 Conjuntos de Dados

M-Lab NDT (Medições de desempenho de rede Internet)

Este dataset possui dados de desempenho de rede (throughput, RTT, perda de pacotes) com acesso público via BigQuery. Para este projeto, será usado um subconjunto.

- O conjunto de dados será fornecido no arquivo **ndt_tests_tratado.csv**.
- As variáveis incluem: **throughput** de download e upload (em bits por segundo), **RTT** de download e upload (em segundos) e **fração de perda de pacotes** (percentual).
- O dataset possui 13 clientes e 7 servidores. A primeira coluna contém a data e a hora da coleta dos dados.

3 Tarefas do Projeto

3.1 Análise Exploratória de Dados (EDA)

- Calcule as estatísticas descritivas: média, mediana, variância, desvio padrão, e **quantis selecionados** (por exemplo, 0,9, 0,99, etc.). **Justifique a escolha** dos quantis mais relevantes para a análise de desempenho de rede (por exemplo, observar a cauda da latência).
- As estatísticas devem ser calculadas para cada variável de interesse (throughput, RTT e fração de perda), **para cada cliente e cada servidor**. Elabore tabelas de resumo e comente as diferenças observadas entre os clientes e entre os servidores.

- Selecione **dois clientes** ou **um cliente e um servidor** que apresentem **comportamentos distintos ou interessantes** para as análises gráficas e modelagem.
- Para os clientes e servidores selecionados, crie os gráficos: histograma, *boxplot*, *scatter plot* (escolha um par de variáveis relevante). Analise e comente as distribuições observadas.
- A partir dos gráficos, **defina um modelo paramétrico candidato** (e.g., Normal, Gamma, Binomial, etc.) para cada uma das cinco variáveis: throughput (up e down), RTT (up e down) e perda.

3.2 Máxima Verossimilhança (MLE)

Utilize o método da Máxima Verossimilhança (MLE) para estimar os parâmetros dos modelos definidos na seção anterior.

- Defina o **Estimador de Máxima Verossimilhança** $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ para cada modelo. Apresente os valores numéricos.
- **Avaliação do Ajuste:** Crie gráficos comparativos para diagnosticar o ajuste do modelo:
    1. Histograma dos dados reais em conjunto com a função densidade/massa de probabilidade do modelo ajustado usando o $\hat{\boldsymbol{\theta}}_{\text{MLE}}$.
    2. *QQ plot* dos dados reais versus quantis teóricos do modelo ajustado.

3.3 Inferência Bayesiana

O objetivo deste item é você usar os modelos que você escolheu no item acima e fazer a inferência Bayesiana. Para a inferência use uma prior conjugada para que a posterior pertença à mesma família. O objetivo é facilitar os cálculos. Informações sobre priors conjugadas podem ser encontradas em Wikipedia e nos capítulos 2 e 3 do livro Bayesian Data Analysis Third edition.

A inferência Bayesiana consiste em três etapas principais:

1. Definir um modelo de verossimilhança (**likelihood**) para os dados observados;
2. Especificar uma distribuição *a priori* (**prior**) para os parâmetros desconhecidos;
3. Calcular a distribuição *a posteriori* (**posterior**) e, a partir dela, a distribuição preditiva (**posterior predictive**) para novas observações.

Tarefas:

- Especifique uma **prior** para cada modelo escolhido. Justifique brevemente os hiperparâmetros iniciais, por exemplo, prior não informativo, fracamente informativo.
- Especifique um modelo para a **likelihood**.
- Calcule os parâmetros da **posterior**.
- **Cálculo Preditivo e Comparação:**
    1. Divida o dataset em **dados de treino** (para calcular a posterior) e **dados de teste** (os dados reais a serem previstos). Você pode usar a proporção 70% dos dados para treino e 30% para teste.

2. Calcule uma previsão para o período de teste utilizando a **posterior predictive**. Defina a equação para a **posterior predictive** e o valor esperado $\mathbb{E}[R_{\text{novo}} \mid \mathbf{r}]$ dos dados previstos.
3. Compare o valor esperado e a variância da **posterior predictive** com a média e varaiância real observadas nos dados de teste.

- **Comparação de Estimativas MLE vs Bayes:** Compare as estimativas pontuais bayesianas (média da *posterior*, $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{r}]$) com as estimativas *MLE* ($\hat{\boldsymbol{\theta}}_{\text{MLE}}$). Discuta as diferenças e o efeito da *prior*.

4 Sugestões de Modelos para os Dados

4.1 O Modelo Normal–Normal para RTT

Para a modelagem do tempo de ida e volta (RTT), frequentemente utiliza-se a distribuição Normal, particularmente para médias agregadas, devido ao Teorema do Limite Central. Este modelo adota uma *prior* conjugada, assumindo que a variância da *likelihood* é conhecida. **Use o valor $\hat{\sigma}^2_{\text{MLE}}$ obtido na Seção 3.2 como a variância σ^2 da likelihood Normal.**

4.1.1 Especificação do Modelo

Assumimos que as observações do RTT, $R = (r_1, r_2, \ldots, r_n)$, são IID (Independentes e Identicamente Distribuídas).

1. **Likelihood (Verossimilhança):** A distribuição dos dados, condicionada à média μ, é Normal com variância σ^2 conhecida.

$$r_i \mid \mu \sim \mathcal{N}(\mu, \sigma^2). \tag{1}$$

2. **Prior (A Priori):** A distribuição a priori para o parâmetro desconhecido μ é também Normal, caracterizada pela média μ_0 e variância τ_0^2.

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2). \tag{2}$$

4.1.2 Distribuição a Posteriori

Dado que a distribuição Normal é a *prior* conjugada para o parâmetro de média de uma *likelihood* Normal com variância conhecida, a distribuição a posteriori, $p(\mu \mid \mathbf{r})$, é também uma distribuição Normal:

$$\mu \mid \mathbf{r} \sim \mathcal{N}(\mu_n, \tau_n^2). \tag{3}$$

Os hiperparâmetros da *posterior* são determinados pela combinação ponderada (pelas precisões) da informação da *prior* e dos dados. Seja $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$ a média amostral.

- **Variância Posterior (τ_n^2):** O inverso da variância (precisão) da *posterior* é a soma das precisões da *prior* e da *likelihood* dos dados.

$$\tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \right)^{-1}. \tag{4}$$

- **Média Posterior (μ_n):** A média é uma média ponderada da média da *prior* e da média amostral $\bar{r}$, com pesos proporcionais às suas respectivas precisões.

$$\mu_n = \tau_n^2 \left(\frac{\mu_0}{\tau_0^2} + \frac{n\,\bar{r}}{\sigma^2} \right). \tag{5}$$

4.1.3 Distribuição Preditiva Posterior

A distribuição preditiva posterior, $p(R_{\text{novo}} \mid \mathbf{r})$, representa a distribuição de um novo RTT, R_{novo}, dadas as observações passadas $\mathbf{r}$.

A marginal preditiva para R_{novo} é também uma **Distribuição Normal**:

$$R_{\text{novo}} \mid \mathbf{r} \sim \mathcal{N}\big(\mu_n,\ \sigma^2 + \tau_n^2\big). \tag{6}$$

A média da preditiva é simplesmente a média da *posterior* μ_n. A variância preditiva, no entanto, é a soma de duas componentes, refletindo o **princípio de variância preditiva**:

- **Variância de Erro (σ^2):** A incerteza inerente ao processo de geração dos dados (incerteza da *likelihood*).
- **Variância de Estimação (τ_n^2):** A incerteza residual sobre o parâmetro μ após a observação dos dados (incerteza da *posterior*).

A soma dessas variâncias $\sigma^2 + \tau_n^2$ incorpora toda a incerteza no processo de previsão.

4.2 O Modelo Beta–Binomial para a fração de perda

A fração de perda de pacotes (p) em redes de comunicação é uma proporção que se encontra no intervalo $[0, 1]$. Esta característica torna o modelo Binomial uma escolha natural para a *likelihood*, com a distribuição Beta servindo como a *prior* conjugada. O par **Beta–Binomial** é ideal para inferência Bayesiana sobre proporções.

4.2.1 Especificação do Modelo

O modelo é definido pela *likelihood* Binomial e pela *prior* Beta para a probabilidade de perda.

4.2.2 Especificação do Modelo

1. **Likelihood (Verossimilhança):** Considere n_t pacotes enviados e x_t pacotes perdidos no período t. A contagem de perdas é modelada por uma distribuição Binomial, onde p é a probabilidade de perda em uma transmissão:

$$X_t \mid p \sim \text{Binomial}(n_t, p). \tag{7}$$

Observação: Como o dataset fornece a **fração de perda** (percentual), e o número real de pacotes (n_t) transmitidos em um teste NDT varia com o throughput e a duração, você deve **assumir um número fixo de pacotes transmitidos (n_t)** para converter a fração em contagem (X_t), permitindo o uso do modelo Binomial.

Sugestão de Valor a ser Fixado: Para simplificar a análise, sugere-se assumir um número fixo de pacotes por observação, como, por exemplo, $\boldsymbol{n_t = 1000}$. O valor de n_t (ou n_{tot}, o agregado) deve ser apresentado na seção de Inferência Bayesiana.

2. **Prior (A Priori):** A distribuição a priori para a probabilidade de perda $p \in [0, 1]$ é a distribuição Beta, caracterizada pelos hiperparâmetros a_0 e b_0.

$$p \sim \text{Beta}(a_0, b_0). \tag{8}$$

4.2.3 Distribuição a Posteriori

Com um conjunto de dados de treino D (agregados sobre n períodos), seja $x_{\text{tot}} = \sum_{t=1}^{n} x_t$ o número total de perdas e $n_{\text{tot}} = \sum_{t=1}^{n} n_t$ o número total de pacotes enviados.

Devido à conjugação, a distribuição a posteriori para a fração de perda p é também uma **Distribuição Beta**:

$$p \mid D \sim \text{Beta}(a_n, b_n), \tag{9}$$

onde os hiperparâmetros atualizados são:

$$a_n = a_0 + x_{\text{tot}} \quad \text{(Forma } a \text{ atualizada com o total de "sucessos" - perdas)} \tag{10}$$

$$b_n = b_0 + (n_{\text{tot}} - x_{\text{tot}}) \quad \text{(Forma } b \text{ atualizada com o total de "falhas")} \tag{11}$$

4.2.4 Distribuição Preditiva Posterior

A distribuição preditiva posterior, $P(X_{\text{novo}} \mid D)$, é a probabilidade de observar k perdas em um novo período com $n_\star$ pacotes. Ela é obtida marginalizando a *likelihood* Binomial sobre a *posterior* Beta.

A marginal preditiva é a **Distribuição Beta–Binomial**:

$$P(X_{\text{novo}} = k \mid D) = \binom{n_\star}{k} \frac{\text{B}(a_n + k,\, b_n + n_\star - k)}{\text{B}(a_n, b_n)}, \qquad k = 0, 1, \ldots, n_\star, \tag{12}$$

onde $\text{B}(a, b)$ é a função Beta, definida como $\text{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

4.2.5 Média e Variância Preditivas

As estatísticas preditivas para o número de perdas (X_{novo}) e para a fração de perda ($X_{\text{novo}}/n_\star$) são dadas por:

1. **Média Preditiva (Contagem de Perdas):**

$$\mathbb{E}[X_{\text{novo}} \mid D] = n_\star \frac{a_n}{a_n + b_n}. \tag{13}$$

2. **Variância Preditiva (Contagem de Perdas):**

$$\text{Var}[X_{\text{novo}} \mid D] = n_\star \frac{a_n b_n (a_n + b_n + n_\star)}{(a_n + b_n)^2 (a_n + b_n + 1)}. \tag{14}$$

3. **Proporção Preditiva (Fração de Perda):**

$$\mathbb{E}[X_{\text{novo}}/n_\star \mid D] = \frac{a_n}{a_n + b_n}. \tag{15}$$

O termo $\frac{a_n}{a_n+b_n}$ corresponde à média da distribuição Beta a posteriori, que serve como o estimador Bayesiano para a probabilidade de perda p.

4.3 O Modelo Gama–Gama para a Throughput

O throughput (taxa de download ou upload) é uma variável estritamente positiva e frequentemente apresenta assimetria positiva. A distribuição **Gamma** é uma escolha apropriada para modelar tais variáveis. Neste modelo Bayesiano, assumimos uma *likelihood* Gamma com o parâmetro de *shape* (k) conhecido e uma *prior* Gamma conjugada no parâmetro de *rate* (β). **Use o valor $\hat{k}_{\mathbf{MLE}}$ obtido na Seção 3.2 como o parâmetro fixo da likelihood Gamma.**

4.3.1 Especificação do Modelo

1. **Likelihood (Verossimilhança):** As observações de throughput $Y = (y_1, \ldots, y_n)$ são modeladas por uma distribuição Gamma com *shape* $k > 0$ (fixo e conhecido) e *rate* $\beta > 0$ (desconhecido).

$$y_i \mid \beta \sim \text{Gamma}(k, \beta). \tag{16}$$

   A função densidade de probabilidade (PDF) é:

$$f(y \mid k, \beta) = \frac{\beta^k}{\Gamma(k)} y^{k-1} e^{-\beta y}, \quad y > 0. \tag{17}$$

2. **Prior (A Priori):** Escolhemos a *prior* conjugada para o parâmetro de taxa β, que é também uma distribuição Gamma com hiperparâmetros a_0 e b_0.

$$\beta \sim \text{Gamma}(a_0, b_0). \tag{18}$$

4.3.2 Distribuição a Posteriori

A distribuição a posteriori, $p(\beta \mid \mathbf{y})$, pertence à mesma família da *prior* (conjugação), sendo uma **Distribuição Gamma** com parâmetros atualizados a_n e b_n.

$$\beta \mid \mathbf{y} \sim \text{Gamma}(a_n, b_n), \tag{19}$$

onde os novos hiperparâmetros são calculados a partir da soma dos parâmetros da *prior* e das estatísticas dos dados:

$$a_n = a_0 + n\,k, \tag{20}$$

$$b_n = b_0 + \sum_{i=1}^{n} y_i. \tag{21}$$

As propriedades da *posterior* são:

$$\mathbb{E}[\beta \mid \mathbf{y}] = \frac{a_n}{b_n}, \qquad \text{Var}[\beta \mid \mathbf{y}] = \frac{a_n}{b_n^2}. \tag{22}$$

4.3.3 Distribuição Preditiva Posterior

A distribuição preditiva posterior para uma nova observação de throughput Y_{novo} (com *shape* k fixo) é obtida marginalizando a *likelihood* pela *posterior*. O resultado é a **Distribuição Beta-prime Escalada** (*Beta distribution of the second kind*), BetaPrime(a_n, k) escalada por $1/b_n$.

$$p(y_{\text{novo}} \mid \mathbf{y}) = \frac{\Gamma(k + a_n)}{\Gamma(k)\,\Gamma(a_n)} \frac{1}{b_n} \left(\frac{y_{\text{novo}}}{b_n}\right)^{k-1} \left(1 + \frac{y_{\text{novo}}}{b_n}\right)^{-(k+a_n)}, \quad y_{\text{novo}} > 0. \tag{23}$$

As propriedades da distribuição preditiva são:

- **Média Preditiva:** A média existe se $a_n > 1$.

$$\mathbb{E}[Y_{\text{novo}} \mid \mathbf{y}] = \frac{k\, b_n}{a_n - 1} \quad (\text{para } a_n > 1). \tag{24}$$

- **Variância Preditiva:** A variância existe se $a_n > 2$.

$$\text{Var}[Y_{\text{novo}} \mid \mathbf{y}] = \frac{k\,(k + a_n - 1)\, b_n^2}{(a_n - 1)^2\,(a_n - 2)} \quad (\text{para } a_n > 2). \tag{25}$$

5 Estrutura Básica do Relatório

1. **Introdução**: Objetivo do trabalho, análises a serem realizadas e hipóteses iniciais.
2. **Descrição do Dataset**: Variáveis, unidades, intervalo de coleta e pré-processamento.
3. **Análise Exploratória (EDA)**: Estatísticas descritivas (com comentário sobre os quantis), gráficos essenciais e a justificativa para a escolha do modelo paramétrico candidato para cada variável.
4. **Modelagem e Inferência**:
   - **MLE**: Parametrização dos modelos, apresentação dos $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ e avaliação do ajuste com gráficos diagnósticos.
   - **Bayesiana**: Especificação e justificativa da *prior*, cálculo dos parâmetros da *posterior* ($\boldsymbol{\theta}_n$) e apresentação da distribuição preditiva (fórmula, valor esperado e variância).
5. **Discussão e Conclusão**: Interpretação dos resultados, comparação das estimativas MLE vs Bayes (discutindo o impacto da *prior*), comparação da previsão Bayesiana com os dados de teste, e limitações dos modelos aplicados.
6. **Uso de Ferramentas de IA e Link para o código**: Descreva como as ferramentas de IA ajudaram na elaboração do projeto e forneça o link para o código.