

## PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

**Disciplina:** Supervised Machine Learning: Modelos para Dados de Contagem I

**Data:** 14/09/2021

**Rubelmar Maia De Azevedo Cruz Neto**

Professor, como determinar a distribuição que representa melhor os dados? Pensando nas várias distribuições possíveis como, por exemplo, a distribuição de Weibull .

Rubelmar, esta definição é feita com base na experiência do pesquisador, na literatura de referência e com base nos testes a serem feitos. As distribuições teóricas são pensadas de modo a abarcar fenômenos do mundo real, mas nem sempre é fácil adequar à realidade aos modelos propostos.

**William Henrique Stenico**

Mosquei, de onde veio o Theta e o Delta?

William, Delta é a taxa de decaimento e Theta é o parâmetro de forma da distribuição. São parâmetros caracterizadores da forma da distribuição.

**Heitor Rossi Ambiel Facini**

A partir do Lambda, como eu acho a probabilidade dele?

Heitor, o Lambda é o valor de abcissa que nos retorna o valor máximo da função. A probabilidade desse evento será o ponto correspondente no eixo Y.

**Ellen Nogueira**

Boa noite! Como faríamos a ponderação em relação ao número da staff? Porque nesse caso não daria para fazer ponderado (violaç/staff) pq deixaria de ser discreto e daí não daria mais p/ usar Poisson

Ellen, a variável “staff” é numérica, quantidade de pessoas. Nesse caso, não necessita de uma ponderação para ser utilizada como variável explicativa. A utilização ou não da distribuição poisson deve ser verificada em relação à variável dependente, no caso do exemplo do professor fávero, quantidade de violações às regras de trânsito pelos membros da representação diplomática.

**Thiago Silva Neves de Barros**

Se fosse realizada a dumização da variável post, mesmo assim o resultado seria o mesmo?

Thiago, nesse caso em específico sim. Em virtude de ser uma variável com duas categorias e sem categorias não significativas.

**Douglas Pasquali Pedroso**

Todos (cada corpo diplomático) tinham carro(s) pessoal?

Douglas, se compreendi bem a sua pergunta, está questionando sobre a quantidade de carros dentro da representação diplomática. Pode ser que haja representações que tenham mais ou menos carros. Seu raciocínio está correto, embora a quantidade de pessoal seja utilizada como uma espécie de proxy para o “tamanho” da representação, a quantidade de carros talvez fosse uma boa opção para a modelagem também. Caso ainda tenha dúvidas, peço que acione a monitoria por e-mail.

### Alexandre Cesar Vereza Medeiros

Prof. Se os golfinhos nas quadrículas são podem ser vistas como clusters, como analisar eles em movimento, considerando cada "foto" sendo tirada a cada 3 segundos (por exemplo), já que eles se movem?

Alexandre, no caso, o cluster seria a fotografia de dado momento. Uma análise de determinado quadro e não abarcaria a movimentação dos golfinhos.

### Pedro Damasco Silveira

Boa Noite, quando pensamos em uma Isocrona, podemos usar semanas como unidade de medida, exemplo x clientes semana 1, y clientes semana 2?

Sim, Pedro. A princípio, é um exemplo de fenômeno que se ajusta ao escopo da técnica.

### Ariston Farias

Boa noite, como é definida as distâncias das quadrículas? Já que são vários apartamentos.

Ariston, acaba sendo uma decisão do pesquisador, baseado na literatura, na expertise, na sua vivência. Talvez o ponto médio dentro de determinada quadrícula seja uma boa opção para essa análise.

### Carlos Henrique de Oliveira

Professor, sabemos que existem muitas PDFs, por exemplo: Normal, Exponencial, Gama, Weibull, etc... como certamente não poderemos tratar tudo isto no curso. Existe alguma sugestão p/ enquadrar?

Carlos, esta definição é feita com base na experiência do pesquisador, na literatura de referência e com base nos testes a serem feitos. As distribuições teóricas são pensadas de modo a abarcar fenômenos do mundo real, mas nem sempre é fácil adequar à realidade aos modelos propostos.

### Israel Luiz Harmendani Diniz

Esse tipo de algoritmo pode ser usado para previsões de demanda / preço e outras variáveis de forma geoespacial por meio de coordenadas?

Israel, caso as suposições acerca da variável dependente e do fenômeno sejam satisfeitas, a princípio, seria possível sim.

### ANTONIO THYRSO CORSINO PEREIRA de SOUZA

Isocronas (olhando para o tempo) são equivalentes a curvas de nível (olhando para altitude)?

Antônio, no sentido de serem uma delimitação de uma determinada grandeza matemática, creio que sim. São determinadas faixas em que enquadrados alguns valores.

### Wagner Seizo Hokama

Quantidade por tempo não seria frequência de uma variável quali?

Wágner, quantidade por tempo é uma variável quanti. Dez carros por segundo em determinado local, 20 litros de água por segundo, e assim por diante.

**Erick Alexandre Nimtz**

Professor, qual a diferença entre o OLS que estudamos (Ordinary Least Squares) e PLS (Partial Least Squares)? Vamos estudar PLS em algum modelo?

O modelo OLS (acrônimo em inglês para Mínimos Quadrados Ordinários - MQO) é um tipo de método com o objetivo de estimar os parâmetros desconhecidos em um modelo de regressão linear a partir da minimização dos quadrados dos resíduos. Já a chamada Regressão por Mínimos Quadrados Parciais (PLS, do inglês "Partial Least Squares") é uma técnica que tem como objetivo reduzir os preditores a um conjunto menor de componentes (não correlacionados), efetuando a regressão de mínimos quadrados para esses componentes no lugar das variáveis originais. A regressão PLS se mostra útil particularmente quando os preditores são altamente correlacionados ou quando existem mais preditores do que observações e a regressão de mínimos quadrados ordinários gera coeficientes com erros padrão altos.

**Ricardo Logli Matsumura**

não sei se o Professor ainda irá apresentar isso, mas existe algum gráfico que pode ilustrar a Super dispersão quando ela é identificada? Relacionando a dispersão de pontos da regr. Linear.

Ricardo, os histogramas apresentados nos dão uma ideia de como as distribuições se comportam. No entanto, apesar da verificação gráfica ser um indicativo da ocorrência do fenômeno da superdispersão, esta questão só é confirmada com o teste estatístico apropriado, nesse caso o teste de Cameron e Trivedi.

**Leandro Modesto Prates Beltrão**

A taxa média, ou lambda, deveria ser o pico dos histogramas? Se sim, pq os gráficos do R não ficaram assim?

Leandro, se entendi bem sua dúvida, o plot gerado em aula tem três gráficos distintos para lambdas distintos, cada um com um pico correspondente.

**Helder Luiz Fonseca Sant' Ana**

Usando o exemplo dos quadrículos, a observação poderia ser uma data? Quantos eventos ocorreram no dia 1, dia 2, etc?

Sim, Helder. A princípio parece um evento que se encaixa nos pressupostos para utilização da técnica.