

Classificador Bayesiano Ingênuo para o Acento do PB

Bruno Ferrari Guide

Orientador: Marcelo Barra Ferreira
Departamento de Linguística - FFLCH - USP

17 de maio de 2016

Introdução

- ▶ Tópicos dessa apresentação:
 - ▶ Objetivos
 - ▶ Sobre o acento
 - ▶ O Classificador Bayesiano Ingênuo
 - ▶ Resultados
 - ▶ Perspectivas

Objetivos

- ▶ A partir da criação de modelos probabilísticos, eu pretendo apresentar uma discussão sobre o comportamento do acento no PB.
- ▶ Os modelos são baseados em corpus e podem trazer à tona algumas características quantitativas sobre esse comportamento.
- ▶ Os modelos retornam as probabilidades de uma determinada palavra pertencer a alguma categoria acentual (Oxítona, Paroxítona, Proparoxítona) e a partir disso é possível discutir os erros e os acertos de um modelo.

Sobre o acento

Sobre o Acento - 2 tendências

- ▶ O acento no português brasileiro (quase) sempre ocupa uma das três últimas posições da palavra, criando as três categorias acentuais: Oxítona, Paroxítona, Proparoxítona.
- ▶ Duas tendências dão conta da maioria das palavras do PB:
 - ▶ Caso a sílaba final seja pesada, a palavra é oxítona.
 - ▶ Caso a sílaba final seja leve, a palavra é paroxítona.

2 tendências?

- ▶ Problemas com palavras oxítonas terminadas em sílaba leve, como *caqui*, *urubu*
- ▶ Problemas com paroxítonas terminadas em sílaba pesada, como em *mártir*, *câncer*, *difícil*.
- ▶ Problemas com as proparoxítonas de modo geral.
- ▶ O acento é regular, porém tem irregularidades.
- ▶ O acento é irregular, porém tem regularidades.

Sobre modelos probabilísticos

- ▶ Modelo é uma representação formal de um objeto.
- ▶ As vezes, o objeto possui comportamento imprevisível.
- ▶ Na matemática, a área que lida com a imprevisibilidade (ou seja, que a quantifica e formaliza) é a probabilidade.
- ▶ Existem muitas formas de tentar formalizar essa imprevisibilidade, cada uma possui suas vantagens e desvantagens.

O classificador Bayesiano Ingênuo

Classificador

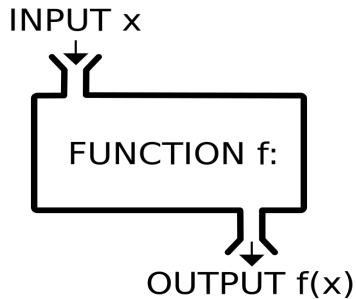


Figura: Um classificador é uma função

Classificador

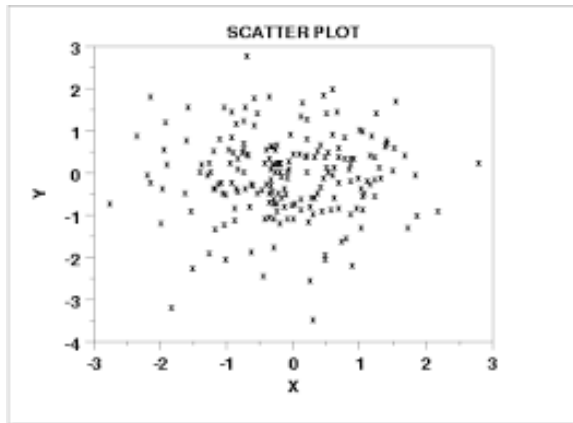


Figura: Em que temos observações como input

Classificador

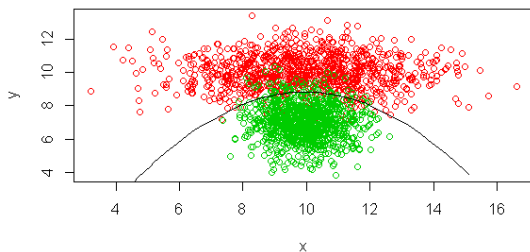


Figura: E as classificamos de acordo com o modelo

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Figura: O teorema (ou regra) de Bayes

CBI - vetor de características

- ▶ Para atribuir a probabilidade de uma palavra (p) pertencer a uma categoria (c), ou seja, para calcular $P(c|p)$, o modelo trata cada palavra como um vetor de traços (\vec{p}).
- ▶ Esses traços são as variáveis observadas.
- ▶ Cada valor possível de cada traço tem uma probabilidade relacionada a cada categoria possível.
- ▶ Essas probabilidades são extraídas do corpus.

Ingênuo

- ▶ Ingênuo pois assume que as diversas características observadas não possuem relação nenhuma entre si.
- ▶ Essa não é uma afirmação necessariamente verdadeira, mas...
- ▶ Isso simplifica o modelo em termos de implementação e computação.

CBI - formalização

- Usando a regra de bayes para calcular $P(c|\bar{p})$, temos que:

1
$$P(c|\bar{p}) = \frac{P(\bar{p}|c) \times P(c)}{P(\bar{p})}$$

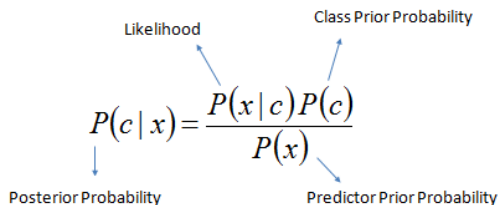
- 2 E o classificador funciona da seguinte maneira: ele me retorna a classe c do conjunto de classes possíveis C que maximiza essa probabilidade.

►
$$CBI(\bar{p}) = \operatorname{argmax} P(c|\bar{p}) : c \in C$$

- 3 Isso resulta em uma simplificação da regra de Bayes, pois toda vez que uma classe c maximize a função em 1, ele irá maximizar a função em 4.

4
$$CBI(\bar{p}) = \operatorname{argmax} (P(\bar{p}|c) \times P(c))$$

CBI - formalização


$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Labels and arrows in the diagram:

- Likelihood (points to $P(x|c)$)
- Class Prior Probability (points to $P(c)$)
- Posterior Probability (points to $P(c|x)$)
- Predictor Prior Probability (points to $P(x)$)

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Figura: Probabilidade de um vetor é igual ao produto da probabilidade de cada um dos elementos do vetor dada a categoria c em questão.

Exemplo

- ▶ Palavra: 'perna'
- ▶ Vetor de traços: [Categoria morfossintática: 'Nome', peso da sílaba final: 'leve', nível de frequência no corpus: 3]
- ▶ Probabilidades:
 - ▶ Oxítone = $p(\text{oxítone}) * p(\text{'nome'}|\text{oxítone}) * p(\text{'leve'}|\text{oxítone})...$
 - ▶ Paroxítone = $p(\text{paroxítone}) * p(\text{'nome'}|\text{paroxítone}) * p(\text{'leve'}|\text{paroxítone})...$
 - ▶ Proparoxítone = $p(\text{proparoxítone}) * p(\text{'nome'}|\text{proparoxítone}) * p(\text{'leve'}|\text{proparoxítone})...$
- ▶ Repare que os meus *priors* vão favorecer as categorias paroxítone, oxítone e proparoxítone nessa ordem.

Implementação

- ▶ Corpus utilizado: ABG \rightarrow 98.000 palavras.
- ▶ Implementação feita em Python.
- ▶ Foi efetuada uma *Cross-validation* dos resultados.
- ▶ Acertos e erros em comparação com acentuação categórica já presente no corpus.

Resultados

Resultados - CBI

- ▶ P = Peso Silábico
- ▶ L = Nível de frequência
- ▶ C = Categoria Morfossintática
- ▶ E = Estrutura silábica (CV-CV)

Resultados CBI

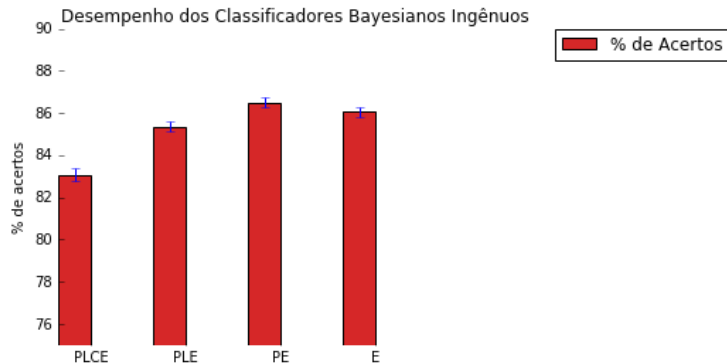


Tabela - CBI

Tabela: Resultados do Classificador Bayesiano Ingênuo

Parâmetros	Média	Desvio
P,L,C,E	83.08	0.33
P.L.E	85.36	0.25
P.E	86.51	0.26
E	86.05	0.26

Resultados - Geral

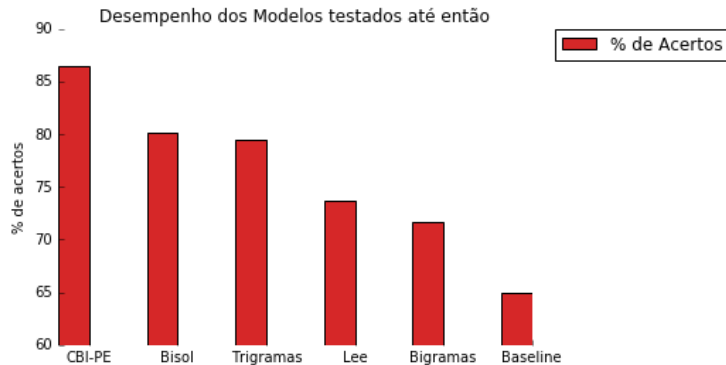


Tabela: Desempenho geral

Tabela: Desempenho Geral

Parâmetros	Média
CBI P.E	86.51
Bisol	80.10
Trigramas	79.40
Lee	73.71
Bigramas	71.69
Baseline	65.59

Perspectivas

Próximos passos

- ▶ Incluir outras versões desse modelo na análise e selecionar as melhores.
- ▶ Escrever, escrever, escrever...

Agradecimento

Muito Obrigado!