# Introdução ao Reconhecimento de Padrões Estáticos

Reconhecimento de Padrões Estatísticos é um termo usado para cobrir todos os estágios de investigação, desde a formulação de um problema e coleta de dados até a descriminação e classificação, avaliação e interpretação de resultados.

Nosso estilo de vida produz uma quantidade muito grande de dados que são coletados a todo o momento, por exemplo: transações bancárias, compras, compras pela internet, performance de trabalho, mensagens digitais e etc. Os dados vêm de forma variada e diversa o que reque alguns procedimentos automatizados para o seu entendimento, assistindo assim os analistas.

Como área de conhecimento o Reconhecimento de Padrões, segundo Webb, A. R. (2011), se desenvolveu significativamente nos anos de 1960, surgindo como um objeto interdisciplinar de estudos. Começaram então a surgir aplicações mais clássicas como reconhecimento de caracteres e diagnóstico médico, atualmente temo outras soluções como a mineração de dados, que atraíram e atraem pesquisas que desenvolvem métodos e aprimoram os já consolidados. Outro ramo de pesquisas é o de desenvolvimento de máquinas com performance “inteligente”, se assemelhando aos seres humanos.

Nessas áreas foram feitos grandes progressos, particularmente nos domínios que interagem com probabilidade e estatística, desenvolvendo novas metodologias e aplicações. Estes ramos são a solidificação de seus predecessores e com o aumento da capacidade de processamento dos computadores progressivo conseguimos superar antigas dificuldades.

O Modelo Básico

Como o tema é uma área interdisciplinar é natural haver contradições com terminologias. Utilizaremos o termo “padrão” para referências a um vetor de dados com *p dimensões* ***x*** *= (x1, ..., xp)T* (*T* significa a transposição do vetor), onde os componentes *xi* são medidas características de um objeto. Por sua vez, as características são variáveis especificadas por um investigador que as considera importantes para a classificação. Durante a discriminação, assume-se que existem *C* grupos ou *classes*, dado que *ω1, ... , ωC* e associados com cada padrão de ***x*** há uma variável categórica *z* representando o pertencimento a um grupo ou *classe*, ou seja, se *z = i*, então o padrão pertence a *ωi,* .

Para a discriminação necessita-se desenhar as especificações de parâmetros para o classificador de padrões, de modo que a resposta seja satisfatória para certa entrada de dados, como mostra a **figura 1.1**. Esta resposta normalmente é uma estimativa da classe à qual o padrão pertence. Assume-se um conjunto de padrões, de uma classe conhecida, (conjunto de treino) usado para desenhar o classificador. Após esta fase podemos estimar o pertencimento de uma classe a um padrão ***x*** para o qual sua identidade é desconhecida.

Chama-se de *indução* o processo de aprender o modelo a partir de *conjunto de treino*, aplicar o modelo treinado para identificar o padrão de uma *classe* desconhecida é o processo de *dedução*.

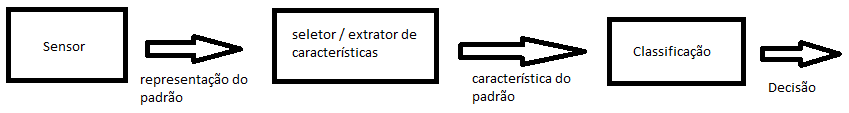


Figura 1.1 Classificador de Padrões

Baseado nisso, um classificador de padrões precisa prover:

* Um modelo descritivo que explique a diferença entre padrões de diferentes classes em termos de características e suas dimensões.
* Um modelo preditivo capaz de identificar um padrão não identificado.

A figura 1.1 simplifica muito o procedimento de classificação de padrões. Os dados podem sofrer vários estágios de transformação antes de estarem preparados para o uso. Essas transformações podem ocorrer na seleção de características, extração de características ou no fim do processo, sempre trabalhando os dados de forma que reduza duas dimensões (números de características), removendo redundâncias ou informações irrelevantes, dando-lhes uma forma mais apropriada para a classificação.

O termo *dimensionalidade intrínseca* se refere ao número mínimo de variáveis necessárias para capturar a estrutura dos dados, este é um processo de *seleção de características*, assumindo a possibilidade de uma combinação não linear das variáveis originais em novas variáveis. A *seleção de características* é o processo de selecionar um subconjunto de um dado conjunto de dados. Em alguns casos não há a possibilidade de uma seleção automática durante a fase de seleção de características, tendo a seleção de características ser feita por um investigador que “conhece” as variáveis que são importantes para a classificação.

Estágios do reconhecimento de padrões

O reconhecimento de padrões consiste de vários estágios que serão enumerados abaixo. Pode ocorrer variações em determinados casos, onde alguns estágios não estarão presentes ou será difícil separar dois estágios.

1. Formulação do Problema: Tem por objetivo entender claramente o objetivo da investigação e planejar os estágios seguintes.
2. Coleta de Dados: realizar a coleta de dados, medindo as variáveis e anotando detalhes sobre o procedimento de coleta.
3. Exame Inicial dos Dados: observar os dados, calcular resumos estatísticos e plotar gráficos para ganhar um melhor entendimento da estrutura dos dados.
4. Seleção de Características ou Extração de Características: Selecionar variáveis apropriadas para a tarefa, podendo ser selecionadas por uma transformação linear ou não dos conjuntos originais (Extração de Características).
5. Classificação de Padrões não Supervisionada (clustering): Pode ser encarado como uma análise exploratória de dados e pode trazer bons resultados para a conclusão do estudo ou pode ser utilizada para viabilizar o processamento de classificação supervisionado.
6. Aplicar Procedimentos de Discriminação ou Regressão: O classificador é criado baseado no conjunto de treino separado.
7. Avaliar os resultados: Aplicar o classificador treinado a um conjunto de teste. A performance do classificador pode ser representada em forma de uma matriz de confusão, onde eij é o número de padrões para a classe ωj que foi classificada como a classe ωi:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Classe verdadeira |  |
| Classe predita |  | ω1 | ω2 | ω3 |
|  | ω1 | e11 | e12 | e13 |
|  | ω2 | e21 | e22 | e23 |
|  | ω3 | e31 | e32 | e33 |

A precisão a é calculada com a seguinte formula:

E a margem de erro é .

1. *Interpretação*: A análise dos resultados pode gerar novas hipóteses que requerem nova coleta de dados, criando assim um ciclo iterativo que só termina quando resolvemos a questão formulada ou quando descobrimos que os dados não são capazes de responder a questão inicial, precisando-se assim ser reformulada.

# Aprendizado e Adaptação

Podemos dizer que todo método que utiliza uma amostragem de treinamento para incorporar informação na construção de um classificador utiliza de aprendizagem, como explicitado acima elas fazem parte dos estágios de classificação, iremos explicar melhor agora o que elas são.

A criação de classificadores consiste em propor uma forma geral de modelo ou classificador, utilizando conjunto de treinamento ou estimando os parâmetros não reconhecidos do modelo. Como a maioria dos problemas interessantes ou práticos são muito complexos, não podemos adivinhar o classificador, por isso temos que considerar a questão da aprendizagem. A aprendizagem se refere a algum algoritmo de gradiente descendente para reduzir os erros em um conjunto de dados de treinamento. Podemos ver a aprendizagem de diversas formas:

* *Aprendizagem Supervisionada*: Nela um “professor” determina a categoria ou o custo para cada padrão dentro do conjunto de treinamento para reduzir o custo sobre estes padrões.
* *Aprendizagem não Supervisionada (clustering)*: Neste caso não há um “professor” e o sistema desenvolve grupos ou “agrupamentos naturais”, sobre os dados analisados. O que é “natural” é definido implicitamente ou explicitamente dentro do grupo. Cada algoritmo desenvolve um agrupamento diferente, dado as características particulares dos conjuntos de padrões ou devido as funções de custo utilizadas.
* *Aprendizagem por Reforço*: O método mais comum de treinar um classificador é tentar computar uma categoria, baseado em uma massa de dados, e se utilizar uma categoria conhecida para aprimorar o classificador. Na aprendizagem por reforço, ou aprendizagem com crítica, não há uma categoria desejada, o único feedback para o aprendizado é se a categoria está certa ou errada, sem dizer o motivo, como se fosse uma crítica.

# Técnicas não Paramétricas

Vimos vários tipos de aprendizado porém não levamos em consideração a densidade do problema. Na aprendizagem supervisionada supomos saber a densidade das funções. Nas técnicas não paramétricas, de aprendizagem não supervisionada, utilizamos uma distribuição arbitrária sem a suposição de que sabemos a densidade do problema.

Há várias formas de técnicas não paramétricas, vejamos algumas:

* Podemos estimar a densidade das funções a partir das amostras disponíveis, se os resultados forem satisfatórios poderemos substituir as densidades verdadeiras pelas achadas.
* Podemos calcular diretamente as probabilidades com . Esse procedimento é o mais próximo das técnicas não paramétricas, que pulam a estimativa das probabilidades para ir direto a tomada de decisão.
* Podemos também transformar o escopo de características na tentativa de aplicar métodos paramétricos dentro do espaço transformado. A isso chamamos de técnicas não paramétricas.

Estimativa de Densidade

Esta estimativa tem por técnica fundamental o fato de que a probabilidade P de um vetor x não pertence a área R é dada por:

(1)

Assim o valor de P é a média da densidade da função p(x), podemos assim estimar a média de p pela estimativa da probabilidade P. Suponha que n amostras x1,...,xn são independes e identicamente distribuídas de acordo com a probabilidade p(x)(Duda). A probabilidade de k de n cair estar contido em R é dado pelo binômio:

(2)

E o valor esperado de K é:

(3)

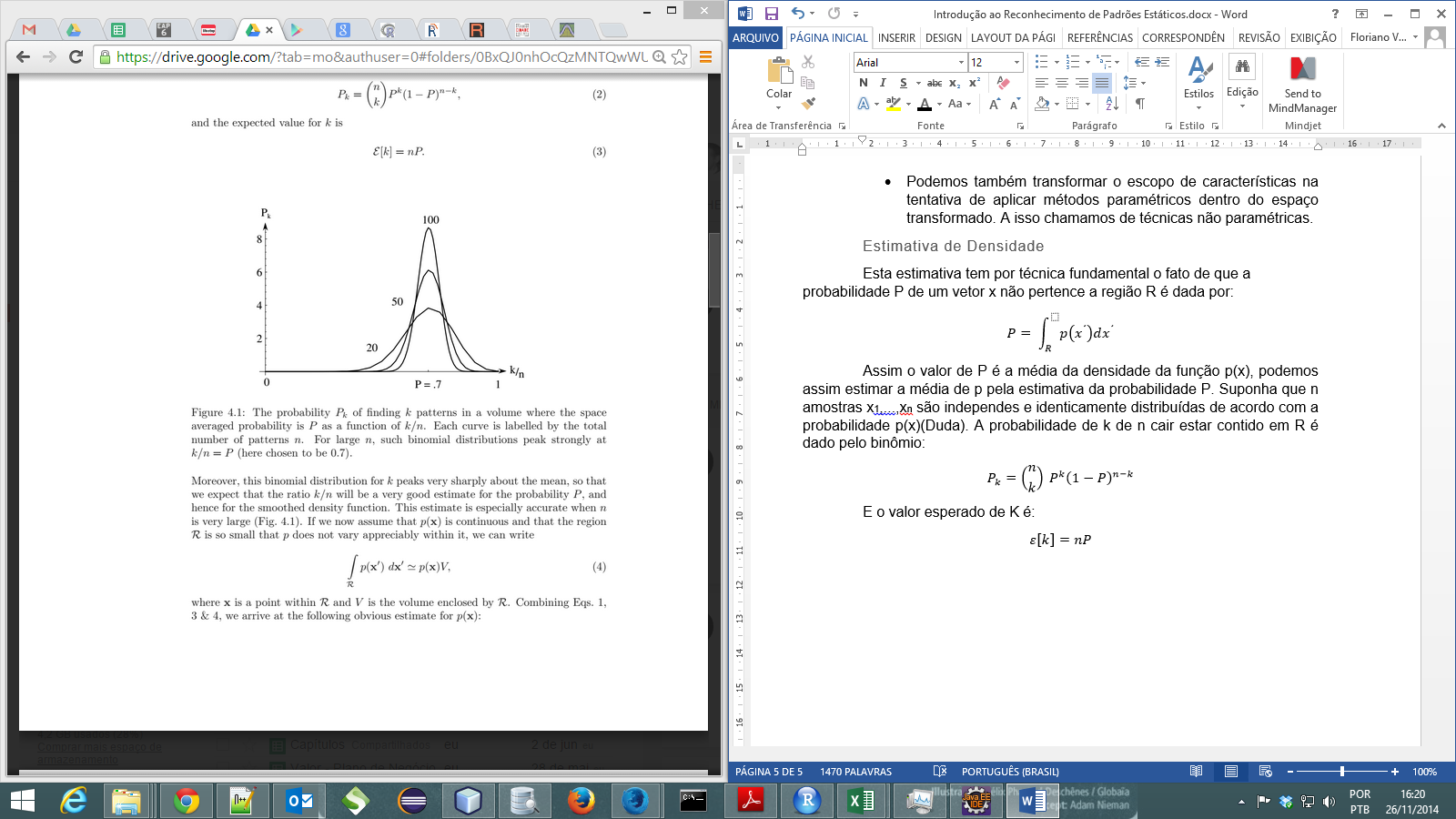


Figura 3.1: Gráfico sobre a probabilidade Pk de encontrar k padrões em um volume onde a área média provável é P como uma função k/n. Cada curva é rotulada pelo total de padrões de n. Para um n grande, como uma distribuição binomial atinge o pico em k/n = P, neste caso 0.7.

Podemos observar que o pico para as distribuições de k são muito próximos da média, por isso esperamos que k/n seja uma boa estimativa para P, e consequentemente, para função de densidade. Assumindo-se que p(x) é continuo e que a área de R é tão pequena que p não varia , podemos supor:

(4)

Onde x é um ponto de R e V é o volume dentro de R.

Combinando as equações 1, 3 e 4 chegamos a:

(5)

Se corrigirmos o volume de V e adicionarmos mais amostras de treinamento, a razão vai provavelmente convergir como o desejado, mas apenas teremos obtido o valor da média da área de p(x).

(6)

Para resgatar o valor de p(x) ao invés de apenas sua média, precisamos aproximar V de zero. Porém se corrigirmos o número de amostras em n e aproximar V de zero, a região R vai ficar tão pequena que não vai mais abrigar nenhuma amostra e a estimativa será inútil.

Do ponto de vista prático, o numero de amostras é sempre limitado, então não podemos permitir que o valor de V fique muito pequeno, sendo assim é necessário admitir certa variância na razão e uma certa densidade da função p(x).

Do ponto de vista teórica, é interessante perguntar como lidar com uma quantidade ilimitada de amostras. Para determinar a densidade em x forma-se uma sequência de regiões R1, R2,..., contendo x, sendo que a primeira região com uma amostra, a segunda com duas e assim por diante. Assim, Vn é o volume de Rn, kn é número de amostras dentro de Rn e pn(x) é enésima estimativa para p(x).

(7)

Se pn(x) é para convergir em p(x), três condições parecem ser necessárias (Duda):

A primeira condição afirma que o espaço médio irá convergir para p(x), dado que as regiões encolhem de forma não uniformemente variada e que p(-) é contínuo em x. A segunda condição somente faz sentido se , assumindo-se que a frequência da razão irá convergir para P. A terceira condição é necessária de dada pela equação 7 for convergir completamente, além disso diz que mesmo com um grande número de amostras eventualmente caiam dentro da menor região de Rn, elas ainda irão formar um pequena fração do total de amostras.

Há duas formas comuns de obter sequências de regiões para satisfazer essas condições. Uma se chama Parzen-window (Janela de Parzen) que consiste em diminuir a região inicial especificando o volume Vn como uma função de n, como , tem se então de verificar se kn e se comportam apropriadamente, ou seja , se pn(x) converge para p(x). No presente estudo não iremos trabalhar com este método, ao em vex disso trabalharemos com kn-nearest-neighbor (kn­ –vizinhos –próximos). Este método consiste em especificar kn como alguma função de n como , fazendo com que o volume de Vn cresça até abarcar kn vizinhos de x.

# Estimação com Kn-Nearest-Neighbor (Kn-vizinhos-mais-próximos)

Esta estimativa consiste em estimar p(X) de n amostras de treinamento das quais podemos centralizar em torno de x até que venha a abranger kn amostras, onde kn é uma função específica de n. Se a densidade é alta, próxima de x a célula será pequena, o que permite uma boa acurácia, se for pequeno a célula irá crescer, porém eventualmente irá atingir as zonas de maior densidade, de qualquer forma podemos supor a seguinte função:

(8)

É desejável que kn seja diretamente proporcional a n quando esta tenda ao infinito, garantindo assim que será uma boa estimativa da probabilidade de um ponto cair em uma célula de voluma Vn. Outra vantagem é garantir que kn cresça suficientemente devagar para que o tamanho da célula necessário para capturar kn seja reduzido a 0.

Esta técnica possui uma estimativa pobre, ele melhora consideravelmente quando n fica cada vez maior, apesar de sua estimativa da integral seja infinita. Isso é compensado pelo fato de pn(x) nunca chega a zero, o que é muito vantajoso em um cenário de muitas dimensões. Vajamos a fórmula de estimativa e um quadro explicativo:

(9)

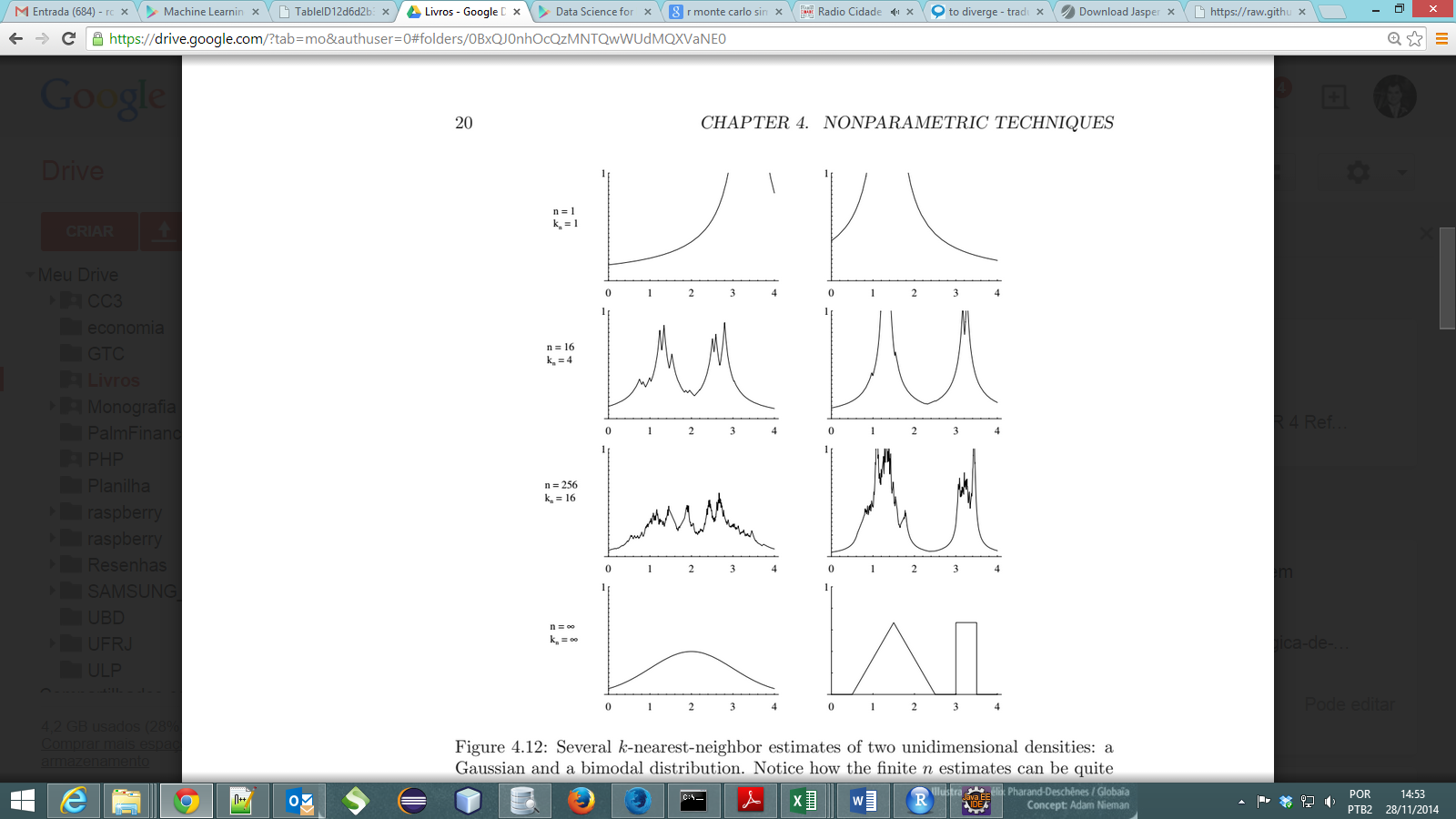


Figura 4.1: Várias estimativas com knn com duas densidades unidimensionais: Uma Gausaniana e outra uma distribuição bimodal.

Podemos utilizar essas técnicas para estimar a posteriori a probabilidade de um conjunto n amostras classificadas usando as amostras para estimar a densidade envolvida. Se suponha que capturemos k amostras ao colocarmos uma célula de volume V em x e que o resultado ki­­ seja classificado por , podemos então estimar com:

(10)

Podemos então utilizar a seguinte fórmula para estimar

(11)

Esta fórmula é uma estimativa a posteriori de que apenas uma fração das amostras são classificadas como . Podemos utilizar como margem de erro selecionamos a categoria mais frequente dentro da célula.

Para decidir o tamanho da célula pode-se expandir V­n até um número específico de amostras sejam capturadas através de uma função, por exemplo, . Conforme n tende ao infinito, um número infinito de amostras se enquadram dentro de uma célula infinitamente pequena (Duda), com isso temos podemos aprender a probabilidades desconhecidas, com uma certa precisão e eventualmente conseguir um performance mais otimizada.

# As Regras do Vizinho mais Próximo

A regra do vizinho mais próximo consiste em classificar ***x*** a partir rótulo mais frequente dentro das *k* amostras mais próximas. Nota-se que se *k*  é fixo e o número de *n* amostras possa se aproximar do infinito, então todos os *k* vizinhos mais próximos vão convergir para ***x***. Portanto, nos casos de um único vizinho próximo, os rótulos em cada um dos *k-vizinhos-mais-próximos* são variáveis aleatórias, as quais assumem valores de forma independente, com probabilidade . A regra dos *k-vizinhos-mais-próximos* seleciona se a maioria dos *k* vizinhos mais próximos são rotulados como , podemos utilizar a seguinte fórmula:

(12)

Sendo assim, quando maior for o valor de *k*, maior é a probabilidade de que será escolhido. Podemos demonstrar que se k é ímpar, uma grande amostra de duas classes possui uma margem de erro definida pela função que define a menor função de maior que o resultado da seguinte fórmula:

(13)

A primeira parte da soma representa a categoria com a probabilidade de erro devido a i pontos vindos da categoria com menor probabilidade e k – i > i pontos vindos de outra categoria. O segundo termo da soma é a probabilidade de k – i pontos virem da categoria com a menor probabilidade e i + 1 < k – i virem da categoria com maior probabilidade. Ambos os casos constituem erros para a regra de *k-vizinhos-mais-próximos*, por isso precisamos adicioná-los para encontrar a probabilidade de erro.

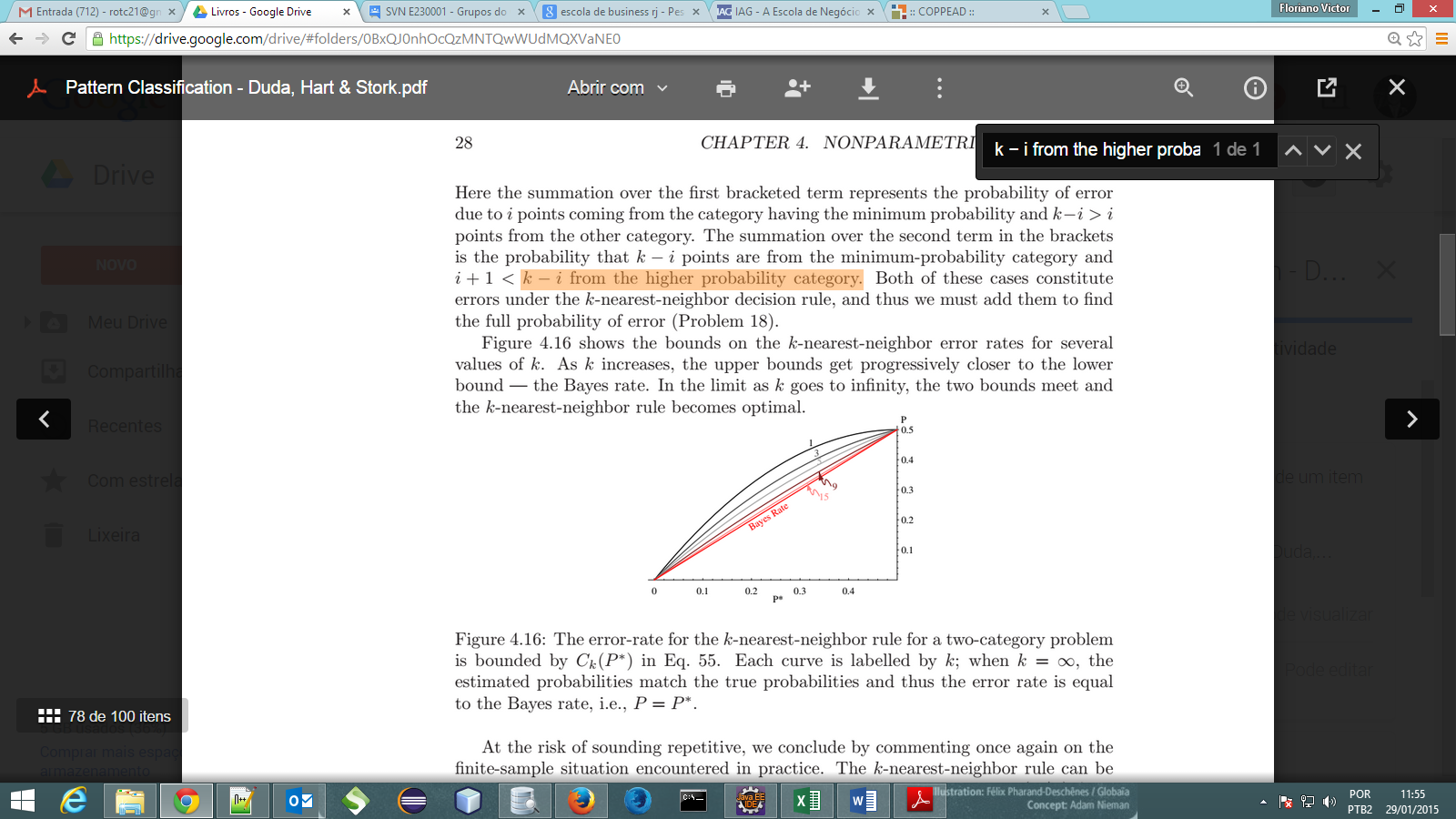


Figura 4.2: A margem de erro para a regra de k-vizinhos-mais-próximos para um problema de duas categorias limitado por na equação 13. Cada curva é rotulada por k, quando k = ∞, a probabilidade estimada corresponde a verdadeiras probabilidades, assim a margem de erro corresponde a P = P­­­\*.

A figura 4.2 mostra como a margem de erro afeta o *k-vizinhos-mais-próximos* para diferentes valores de *k*. Conforme k aumenta, as linhas superiores se aproximam das inferiores, a taxa de erro de Bayes. No limite, conforme k cresce ao infinito as duas curvas se encontram e o *k-vizinhos-mais-próximo* se torna opcional.

Sendo assim, com uma amostra finita, é desejável que *k* tenha um valor grande para se obter uma estimativa a posteriori confiável de . Espera-se que todos os valores de seja próximo de x para que seja bem similar a . Isso força a escolher um valor de K que se apenas um fração do total de amostras. Somente no limite onde n tende ao inifinito que podemos assegurar uma execução otimizada da regra dos *k-vizinhos-mais-próximos.*