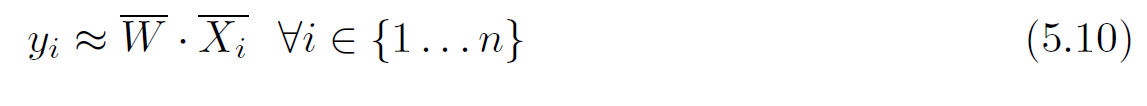


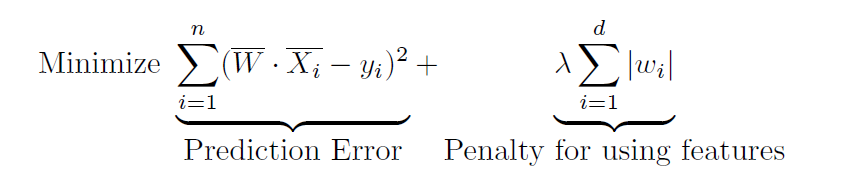
**Aprendizado de máquina para texto**

**5.2.6 Modelos de seleção de recursos incorporados**

Muitos modelos de classificação e regressão fornecem a capacidade de realizar a seleção de recursos incorporados, aproveitando a saída de etapas intermediárias. A seleção de recursos é realizada com o uso de regularização a fim de reduzir o **sobreajuste (overfitting)**, que é semelhante em princípio aos objetivos da seleção de recursos. Como resultado, as saídas intermediárias desses algoritmos regularizados fornecem informações úteis para a seleção de recursos. Por exemplo, considere o seguinte modelo de **regressão linear** em que o dependente numérico variável yi é previsto usando a seguinte relação linear com as variáveis de recursos yi variável yi é previsto usando a seguinte relação linear com as variáveis de recursos i.



**A notação representa um vetor *d-*dimensional de coeficientes que é aprendido pelo modelo de treinamento.** Este vetor é calculado resolvendo o seguinte modelo de otimização:



Aqui, *λ>* 0 é um parâmetro de regularização, que controla a gravidade da penalidade. Tal penalidade garante que a otimização não atribuirá um grande coeficiente diferente de zero para aquele **recurso (*feature*)**, a menos que o recurso transmita informações importantes e insubstituíveis sobre a variável dependente. **A penalização de recursos é conhecida como regularização**. O tipo de penalidade discutido acima é referido como o L1-penalidade(penaty), e tem a notável propriedade de favorecendo um vetor coeficiente em que muitos valores de i são zero. Tais recursos são efetivamente eliminados porque eles não terão nenhuma influência na predição de instâncias de teste. A ideia natural na seleção de recursos embutidos é que ela aproveita os mecanismos embutidos (regularização) por muitos algoritmos para evitar ajustes excessivos. **Afinal, o principal objetivo da seleção de recursos também é a prevenção de ajustes excessivos**.

**5.2.7 Truques (Tricks) de engenharia de recursos**

Dois tipos de truques de engenharia de recursos são comumente usados no domínio do texto.

1. O **primeiro truque é feito para se livrar da dispersão**, o que pode ser um problema para alguns classificadores, como árvores de decisão.
2. A **segunda técnica usa técnicas de mineração de representação** para embutir representações sequenciais de texto em representações multidimensionais. A última abordagem é capaz de alavancar as informações de ordenação sequencial entre as palavras para incorporar um maior conhecimento semântico na aprendizagem. Já que a segunda abordagem será discutida no Cap.10, a seguir irá discutir apenas os métodos de engenharia de recursos usados para lidar com a dispersão.

A escassez pode causar desafios com certos tipos de classificadores, como árvores de decisão, que usam atributos ***um por vez*** no processo de modelagem. **Uma vez que cada termo contém informações relevantes para apenas um pequeno subconjunto de documentos nos quais está presente, e a ausência de termos é uma informação ruidosa, muitas vezes causa excesso de ajuste quando os classificadores tomam decisões importantes com atributos individuais.** Portanto, em tais casos, métodos como **análise semântica latente (LSA)** não são úteis apenas para a redução de dimensionalidade, mas podem ser vistos como métodos de engenharia de recursos que permitem o uso de certos tipos de classificadores. Uma variante particular de LSA, conhecida como *Conjunto de rotação* é particularmente útil para implementações centradas em conjunto. A ideia básica é usar a seguinte abordagem:

Divida aleatoriamente o *d* termos em K subconjuntos disjuntos de tamanho d / K para criar K conjuntos de dados projetados;

Executar LSA em cada conjunto de dados projetados para extrair r << d / K **recursos (features);**

Agrupe todos os recursos extraídos para criar um (K . r)-conjunto de dados dimensionais;

Aplique um classificador na nova representação;

Essa abordagem pode ser aplicada várias vezes e a previsão de uma instância de teste pode ser calculada em várias transformações. Um classificador particularmente comum que é usado com esta abordagem é a *árvore de decisão,* e o classificador resultante é referido como o ***Floresta* *de Rotação (Random Forest).***Outro método de engenharia de recursos é o discriminante linear de Fisher, que fornece *discriminativo* direções no espaço. Esses métodos também foram usados em conjunto com árvores de decisão.

5.3 O Modelo Naive Bayes

O classificador de Bayes usa um modelo generativo probabilístico que é idêntico ao modelo de mistura usado para agrupamento. O modelo assume que o corpus é gerado a partir de uma mistura de diferentes classes. O processo gerador, que é aplicado uma vez para cada documento observado, é o seguinte:

1. Selecione o *rth* classe (componente da mistura) *Cr* com probabilidade anterior *αr =* *P(Cr).*
2. Gere o próximo documento a partir da distribuição de probabilidade para *Cr.* **As escolhas mais comuns são as distribuições Bernoulli e multinomial**.

Os dados observados (treinamento e teste) são considerados resultados desse processo gerador, e os parâmetros desse processo gerador são estimados de modo que a probabilidade logarítmica desse conjunto de dados a ser criado pelo processo gerador seja maximizada. **Geralmente, apenas os dados de treinamento são usados para estimar os parâmetros, porque os dados de treinamento contêm informações adicionais sobre a identidade do componente da mistura que gerou cada documento.** Posteriormente, esses parâmetros são usados para estimar a probabilidade de geração de cada documento de teste não rotulado de cada componente (classe) da mistura. Isso resulta em uma classificação probabilística de **documentos não rotulados.**

Cada cluster *Gr* no algoritmo de maximização de expectativa é análogo a

uma classe *Cr* neste cenário. Pode-se ver o Naıve Bayes como uma simplificação do algoritmo de maximização de expectativa iterativa em que a presença de rótulos permite a execução de a abordagem em uma única iteração. Ao contrário do clustering, o processo de treinamento na classificação usa um único a aplicação de M-steps (em dados rotulados) e a predição probabilística de instâncias de teste uma única aplicação da E-steps nas instâncias de teste não rotuladas (para estimar probabilidades posteriores). Além disso, o classificador de Bayes tem modelos de Bernoulli e multinomiais análogos aos usados em agrupamento.

**5.3.1 O Modelo Bernoulli**

No modelo de Bernoulli, assume-se que apenas a presença ou ausência de cada termo no documento é observada. Portanto, as frequências dos termos são ignoradas e a representação do espaço vetorial de um documento é um vetor binário esparso. O modelo Bernoulli assume que o *j*th termo, *tj,* no léxico está presente em um documento gerado a partir do *r*th classe (componente da mistura) com probabilidade p(r)j. Então, a probabilidade P(|Cr) da geração do documento do componente da mistura Cr é dada pelo [produto](#page4) do *d* diferente Probabilidades de Bernoulli correspondentes à presença ou ausência de vários termos:



Uma suposição importante aqui é que a presença ou ausência dos vários termos são condicionalmente independentes no que diz respeito à escolha da classe. Portanto, pode-se expressar a probabilidade conjunta dos atributos em como o produto dos valores correspondentes em atributos individuais. Essa suposição também é conhecida como suposição de Bayes, que também é a razão pela qual o método é referido como um classificador Naive Bayes. **O termo “ingênuo” (Naive) é usado porque esse tipo de aproximação geralmente não é verdadeiro em ambientes reais.**

A principal tarefa na fase de treinamento do classificador de Bayes é estimar o (máximo verossimilhança) valores das probabilidades anteriores αr e probabilidades gerativas específicas de classe p(r)j.

Esses parâmetros são estimados de forma que os dados observados tenham a probabilidade máxima sendo gerados pelo modelo e, em seguida, usados para realizar a previsão dos rótulos de instâncias de teste invisíveis. Pode-se resumir esse processo da seguinte forma:

* **Fase de treinamento:** Estimar os valores de máxima verossimilhança dos parâmetros p (r) j e αr usando apenas os dados de treinamento.
* **Fase de predição :** Use os valores estimados dos parâmetros para prever o de cada instância de teste não rotulada.

**A fase de treinamento é executada primeiro, seguida pela fase de previsão. Porém, como a fase de predição de um classificador Bayes ingênuo é a chave para entendê-la, apresentaremos a fase de predição antes da fase de treinamento. Portanto, a seção a seguir presumirá que os parâmetros do modelo já foram aprendidos na fase de treinamento**.

**5.3.1.1 Fase de Predição**

A fase de **previsão/predição** usa a regra de Bayes de probabilidades posteriores para prever uma instância. A ideia básica é que o aluno use a frequência agregada de cada classe no treinamento dados para aprender uma anterior probabilidade αr = P(Cr), de cada classe. Posteriormente, ele precisa estimar a posterior probabilidade P(Cr|) depois de observar um específico documento (com binário representação = (z1 . .zd)) para o qual o rótulo não é conhecido. Esta estimativa fornece uma previsão probabilística para a instância de teste de pertencer a uma determinada classe.

De acordo com a regra de Bayes de probabilidades posteriores, a probabilidade posterior de sendo gerado pelo componente da mistura Cr do *r*th classe pode ser estimada como segue:

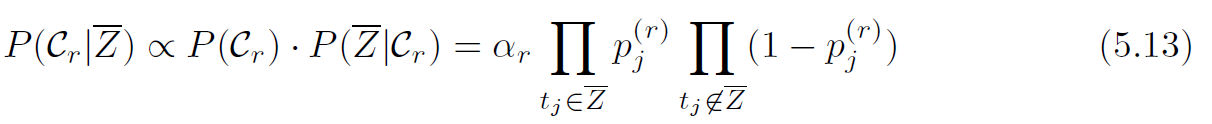


Uma constante de proporcionalidade3 é usado em vez do *P()* no denominador, porque o a probabilidade estimada só é comparada entre várias classes para determinar a classe prevista, e *P()* é independente da classe.

2 Embora i é ~~um~~ vetor binário, estamos tratando-o como um conjunto quando usamos uma notação de associação de conjunto gostar *tj* *∈* i*.* Qualquer vetor binário também pode ser visto como um conjunto de 1s nele.

3 A constante de proporcionalidade pode ser facilmente inferida, garantindo que a soma das probabilidades em todas as classes é 1. Como veremos mais tarde, existem cenários associados a instâncias de classificação para pertencer a classes específicas, onde a constante de proporcionalidade importa.

Uma observação importante aqui é que todos os parâmetros do lado direito da condicional podem ser estimados usando o modelo de Bernoulli. Nós expandimos ainda mais o relacionamento usando a distribuição de Bernoulli do [seguinte](#page4) modo:



Observe que todos os parâmetros do lado direito são estimados durante a fase de treinamento discutida abaixo. Portanto, agora se tem uma probabilidade estimada de cada classe ser prevista até um fator constante de proporcionalidade. A classe com a probabilidade posterior mais alta é prevista como a relevante, embora a saída às vezes seja fornecida na forma de probabilidades. É digno de nota que esta etapa é idêntica à etapa-E (E-step) usada para modelagem de mistura em agrupamento, exceto que é aplicado apenas às instâncias de teste não rotuladas.

**5.3.1.2 Fase de Treinamento**

A fase de treinamento do classificador Bayes usa os dados de treinamento rotulados para estimar os valores de máxima verossimilhança dos parâmetros. É evidente que precisamos estimar dois conjuntos de parâmetros, que são as probabilidades anteriores αr e o gerador Bernoulli parâmetros, p(r)j para cada componente da mistura.

As estatísticas disponíveis para a estimativa de parâmetros incluem o número de documentos rotulados nr pertencentes à *r*th class Cr, e o número, m(r)j, dos documentos pertencentes à classe Cr que contêm o termo tj. O máximo as estimativas de probabilidade destes parâmetros podem ser mostradas como sendo as seguintes:

*1 Estimativa de probabilidades anteriores:* Uma vez que os dados de treinamento contêm nr documentos para o *r*th classe em um tamanho de corpus de n, a estimativa natural para a probabilidade anterior da classe é a seguinte:



Se o tamanho do corpus for pequeno, é útil realizar a suavização **Laplaciana** adicionando um pequeno valor *β>* 0 para o numerador e *β . k* ao denominador:



O valor preciso de *β* contém a quantidade de suavização e geralmente é definido como 1 na prática. Quando a quantidade de dados é muito pequena, isso resulta nas probabilidades anteriores sendo estimadas mais próximas de 1 /*k,* o que é uma suposição sensata na ausência de dados suficientes.

*2 Estimativa de parâmetros de mistura condicionados por classe:* A mistura condicionada por classe parâmetros, p(r)j são estimados da seguinte forma:



É particularmente importante usar o Laplaciano a suavização dos problemas condicionados à classe probabilidades porque um termo particular tj pode nem mesmo está presente no documento de classe *r*th, principalmente quando treinamento do corpus é pequeno. Nesse caso, um estimaria o valor correspondente de p(r)j para 0. Como resultado natural da multiplicação, a presença do termo tj em um documento invisível sempre levará a uma probabilidade estimada de 0 para a classe *r*th. **Essas previsões são frequentemente erradas, e são causados por ajuste excessivo ao pequeno tamanho dos dados de treinamento.**

A suavização Laplaciana da estimativa de probabilidade condicionada por classe é realizada como segue. Deixar da ser o número médio de 1s na representação binária de cada documento de treinamento e d ser o tamanho do léxico. A ideia básica é adicionar um liso Laplaciano parâmetro ing γ> 0 ao numerador da Eq. 5,16 e d γ/dS ao denominador:



**O valor de *γ* geralmente é definido como 1 na prática. Quando a quantidade de dados de treinamento é muito pequeno**, essa escolha leva a um valor padrão de uma/ d para p (r) que reflete o nível de j, esparsidade na coleção de documentos.

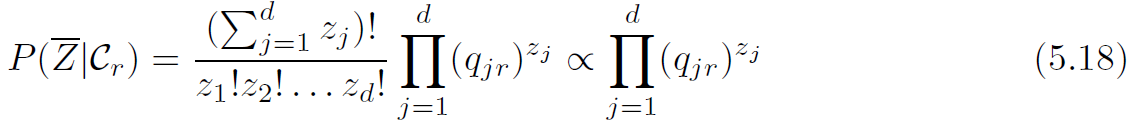
Digno de nota que a fase de treinamento no classificador Bayes é uma variante simplificada da **etapa M(M-step)** usada no modelo de mistura para agrupamento. Esta simplificação é porque etiquetado dados de treinamento estão disponíveis para inferir a associação de documentos em componentes de mistura.

**5.3.2 Modelo Multinomial**

Enquanto o modelo de Bernoulli usa apenas a presença de ausência de termos nos documentos, o modelo multinomial usa explicitamente suas frequências de termo. Assim como o parâmetro p(r)j no modelo de Bernoulli denota a probabilidade de um termo ser observado em um determinado componente, o parâmetro qjr no modelo multinomial denota a presença fracionária de prazo tj no *r*th componente da mistura, incluindo o efeito das repetições. Os valores de qjr soma a 1 para um componente de mistura particular r sobre todos os termos (ou seja, ∑dj=1 qjr=1).

O processo gerador para o modelo de mistura multinomial primeiro seleciona a classe *r*th

(componente da mistura) com probabilidade αr = P (Cr). Em seguida, ele lança um dado carregado (pertencente ao *r*th classe) *L* vezes para gerar um documento com *L* tokens (contagem de repetições). O dado carregado tem tantas faces quanto o número de termos *d*, e a probabilidade do *j*th cara aparecer é dado por qjr para o dado pertencente ao *r*th classe. Portanto, se o dado for lançado *L* vezes, então o número de vezes que cada rosto aparece fornece o número de vezes que cada termo aparece no documento observado. Se assumirmos que o vetor de frequência do documento É dado por (z1. .zd), então a probabilidade gerativa do euo documento é dado pela seguinte distribuição multinomial:



A constante de proporcionalidade mantém-se para fixo e classe variável, porque depende apenas de e é independente da classe Cr.

O processo geral de previsão e treinamento no modelo multinomial é muito semelhante ao do modelo de Bernoulli. Como no caso do modelo Bernoulli, pode-se use a regra de Bayes e a Eq. 5,18 para derivar os seguintes valores de para o posterior estimado pertence à classe Cr:



Se necessário, a constante de proporcionalidade pode ser inferida garantindo que as probabilidades posteriores sobre todas as classes somam 1. A classe com a maior probabilidade posterior pode ser prevista como a relevante para a instância de teste *.*

Para calcular os valores do lado direito da Eq. 5,19, só precisa estimar os parâmetros *αr* e *qjr* durante a fase de treinamento. A presença fracionária de cada classe nos dados de treinamento é usada como a estimativa de *αr.* A suavização laplaciana pode ser usado se necessário. Além disso, se *ν(j,r)* é o número de vezes que o termo *tj* aparece nos documentos pertencentes à classe *r (*com crédito proporcional dado às repetições em um documento único), então a estimativa qjr pode ser calculado da seguinte forma:



Também se pode ver esta estimativa como a fração do número de *tokens (*ou seja, posições) em uma classe que corresponde a um determinado termo. Isso é diferente do modelo de Bernoulli, que estima as probabilidades condicionadas por classe como a fração de documentos específicos de classe contendo um determinado termo. Também é possível usar a **suavização Laplaciana** para suavizar a estimativa. Neste caso, adicionamos um pequeno valor *γ>* 0 para o numerador e *γ · d para o denominador. Isso resulta na seguinte estimativa:*

**

É comum definir *γ* a 1. Este tipo de suavização viesa a estimativa da probabilidade de cada um dos *d* faces na rolagem de dados multinomial em direção a 1 /*d,* o que implica que todos os termos são igualmente favorecidos. Esta é uma suposição razoável na ausência de dados suficientes.

**5.3.3 Observações Práticas**

A suposição ingênua de independência condicional nunca é realmente verdadeira em ambientes práticos. Apesar desse fato, as previsões reais são surpreendentemente robustas. Usar suposições mais complicadas muitas vezes acaba superando os dados.

Levanta-se uma questão natural sobre quando é preferível utilizar o Bernoulli ou o modelos multinomiais.Observe que o modelo de Bernoulli usa a presença e a ausência de termos em um documento, mas não usa o termo frequências. Os dois fatores principais são:

1. o comprimento típico de cada documento e,
2. o tamanho do léxico a partir do qual os termos são extraídos.

**Para documentos curtos que têm uma representação não esparsa em relação a um pequeno léxico, faz sentido usar o modelo de Bernoulli. Em documentos curtos, há um número limitado de repetições de termos, o que reduz o ganho obtido com a inclusão de informações de frequência. Além disso, se o tamanho do léxico for muito pequeno e a representação do espaço vetorial não for esparsa, mesmo a ausência de um termo em um documento é informativa. Quando a representação do documento é esparsa, as informações sobre a ausência de termos são ruidosas, o que prejudica o modelo de Bernoulli. Além disso, ignorar as informações de frequência também aumentará a imprecisão do modelo de Bernoulli. Portanto, faz sentido usar o modelo multinomial nesses casos.**