Charu C. Aggarwal

Aprendizado de máquina para texto

123

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, 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 (consulte a seção6.2.2 do cap. 6), em que o dependente numérico variável y_{eu} é previsto usando a seguinte relação linear com as variáveis de recursos $X_{eu:}$

$$y_{eu} \approx \overline{C} \cdot \overline{X_{eu}} \; \forall eu \in \{1 \ldots n\} \tag{5,10}$$

A notação $\overline{C}$ representa um d-vetor dimensional de coeficientes que é aprendido pelo modelo de treinamento. Este vetor é calculado resolvendo o seguinte modelo de otimização:

$$\text{Minimizar} \underbrace{\sum^{n} (\overline{C} \cdot \overline{X_{eu}} - y_{2\,i)+}}_{\text{Erro de previsão}} \quad \underbrace{\lambda \sum^{d} / C_{eu} /}_{\text{Penalidade por usar recursos}}$$

Aqui, $\lambda > 0$ 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, 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 eu_{1}-pena, e tem a notável propriedade de favorecendo um vetor coeficiente $\overline{C}$ em que muitos valores de C_{eu} são zero. Tais recursos são efetivamente eliminados porque eles não terão nenhuma influência na predição de instâncias de teste de acordo com a Eq. 5,10. 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. Uma discussão detalhada sobre eu_{1}-a regularização é fornecida na Seção 6.2.2 do cap. 6.

5.2.7 Truques de engenharia de recursos

Dois tipos de truques de engenharia de recursos são comumente usados no domínio do texto. O primeiro truque é feito para se livrar da dispersão, o que pode ser um problema para alguns classificadores, como árvores de decisão. 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;
Agrupe todos os recursos extraídos para criar um $(K \cdot r)$ -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 o *árvore de decisão,* e o classificador resultante é referido como o *Floresta de Rotação [*413] Outro método de engenharia de recursos é o discriminante linear de Fisher (cf. Seção6.2.3 do cap. 6), que fornece *discriminativo* direções no espaço. Esses métodos também foram usados em conjunto com árvores de decisão [82]

5.3 O Modelo Na¨ive Bayes

O classificador ingênuo Bayes usa um modelo generativo probabilístico que é idêntico ao modelo de mistura usado para agrupamento (cf. Seção 4,4 do cap. 4) 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 r^a classe (componente da mistura) C_r com probabilidade anterior $\alpha_r = P(C_r)$.

2. Gere o próximo documento a partir da distribuição de probabilidade para C_r. 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 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 G_r no algoritmo de maximização de expectativa de Sect. 4,4 é análogo a uma aula C_r neste cenário. Pode-se ver o ingênuo 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 da etapa M (em dados rotulados) e a predição probabilística de instâncias de teste é uma única aplicação da etapa E nas instâncias de teste não rotuladas (para estimar probabilidades posteriores). Além disso, o classificador ingênuo 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^b termo, t_j, no léxico está presente em um documento gerado a partir do r^a aula (componente da mistura) com probabilidade $p_{(r)j}$. Então, a probabilidade $P(Z/\overline{C_r})$ da geração

do documento $\bar{Z}$ do componente da mistura C_r é dada[2] pelo produto do d diferente Probabilidades de Bernoulli correspondentes à presença ou ausência de vários termos:

$$P(\bar{Z}_r) = C \prod_{t_j \in \bar{Z}} p_j^{(r)} \prod_{t_j \in \bar{Z}} (1 - p_{(r)}) \quad (5,11)$$

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 $\bar{Z}$ como o produto dos valores correspondentes em atributos individuais. Essa suposição também é conhecida como *suposição ingênua de Bayes,* que também é a razão pela qual o método é referido como um *classificador ingênuo Bayes.* O termo "ingênuo" é usado porque esse tipo de aproximação geralmente não é verdadeiro em ambientes reais.

A principal tarefa no *fase de treinamento* do classificador de Bayes é estimar o (máximo verossimilhança) valores das probabilidades anteriores α_r e probabilidades gerativas específicas de classe $p_{(r)j}$. Esses parâmetros são estimados de forma que os dados observados tenham a probabilidade máxima de 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 α_r usando apenas os dados de treinamento.
- **Fase de previsão:** Use os valores estimados dos parâmetros para prever o de cada classe 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 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 um *anterior* probabilidade $\alpha_r = P(C_r)$, de cada classe. Posteriormente, ele precisa estimar o *posterior* probabilidade $P(C_r|\bar{Z})$ depois de observar um *específico* documento (com binário representação $\bar{Z} = (z_1 \ldots z_d)$) para o qual o rótulo não é conhecido. Esta estimativa fornece uma previsão probabilística para a instância de teste $\bar{Z}$ de pertencer a uma determinada classe.

De acordo com a regra de Bayes de probabilidades posteriores, a probabilidade posterior de $\bar{Z}$ sendo gerado pelo componente da mistura C_r do ra classe pode ser estimada como segue:

$$P(C_r|\bar{Z}) = \frac{P(C_r) \cdot P(\bar{Z}|C_r)}{P(\bar{Z})} \propto P(C_r) \cdot P(\bar{Z}|C_r) \quad (5,12)$$

Uma constante de proporcionalidade[3] é usado em vez do $P(\bar{Z})$ no denominador, porque o a probabilidade estimada só é comparada entre várias classes para determinar a classe prevista, e $P(\bar{Z})$ é independente da classe.

---

[2] Embora $X_{\bar{u}}$ é um vetor binário, estamos tratando-o como um conjunto quando usamos uma notação de associação de conjunto gostar $t_j \in X_{\bar{u}}$. 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 dos 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 na Eq. 5,12 usando a distribuição de Bernoulli da Eq. 5,11 do seguinte modo:

$$P(C_r|\overline{Z}) \propto P(C_r) \cdot P(Z|C_r) = \alpha_r \prod_{t_j \in Z} p_j^{(r)} \prod_{t_j \in Z} (1 - p_j^{(r)}) \quad (5,13)$$

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 usada para modelagem de mistura em agrupamento (cf. Seção4.4.1), 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 na Eq. 5,13. É evidente que precisamos estimar dois conjuntos de parâmetros, que são as probabilidades anteriores α_r e o gerador Bernoulli parâmetros, $p_j^{(r)}$ para cada componente da mistura. As estatísticas disponíveis para o parâmetro estiinformações incluem o número de documentos rotulados n_r pertencendo ao r^a aula C_r, e a número, $m_j^{(r)}$ dos documentos pertencentes à classe C_r que contém o termo t_j. O máximo as estimativas de probabilidade desses parâmetros podem ser as seguintes:

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

$$\alpha_r = \frac{n_r}{n} \quad (5,14)$$

Se o tamanho do corpus for pequeno, é útil realizar a suavização Laplaciana adicionando um pequeno valor $\beta > 0$ para o numerador e $\beta \cdot k$ ao denominador:

$$\alpha_r = \frac{n_r + \beta}{n + k\beta} \quad (5,15)$$

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, $p_j^{(r)}$ são estimados da seguinte forma:

$$p_j^{(r)} = \frac{m_j^{(r)}}{n_r} \quad (5,16)$$

É particularmente importante usar o Laplaciano a suavização dos problemas condicionados à classe habilidades porque um termo particular t_j pode nem mesmo está presente no documento de umentos do ra aula, principalmente quando treinamento; o corpus é pequeno. Nesse caso, um estimaria o valor correspondente de $p_j^{(r)}$ a 0. Como resultado do multiplicativo

natureza da Eq. 5,13, a presença do termo t_j em um documento invisível sempre levará a uma probabilidade estimada de 0 para o r^a classe. 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 baixos. Deixar d_{uma} 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 $y > 0$ ao numerador da Eq. 5,16 e $d\,y / d_{uma}$ ao denominador:

$$p_j^{(r)} = \frac{m_j^{(r)} + y}{n_r + d\,y / d_{uma}} \tag{5,17}$$

O valor de y 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 d_{uma}/d para $p_j^{(r)}$ que reflete o nível de esparsidade na coleção de documentos.

É digno de nota que a fase de treinamento no classificador Bayes é uma variante simplificada da etapa M usada no modelo de mistura para agrupamento (cf. Seção 4.4.1) 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 o modelo multinomial usa explicitamente suas frequências de termo. Assim como o parâmetro $p_j^{(r)}$ no modelo de Bernoulli denota a probabilidade de um termo ser observado em um determinado componente, o parâmetro q_{jr} no modelo multinomial denota a presença fracionária de prazo t_j no ro componente da mistura, incluindo o efeito das repetições. Os valores de q_{jr} soma a 1 para um componente de mistura particular r sobre todos os termos (ou seja, $\sum_{j=1}^{d} q_{jr} = 1$).

O processo gerador para o modelo de mistura multinomial primeiro seleciona o r^a aula (componente da mistura) com probabilidade $\alpha_r = P(C_r)$. Em seguida, ele lança um dado carregado (pertencente ao r^a aula) eu vezes para gerar um documento com eu *tokens* (contagem de repetições). o dado carregado tem tantas faces quanto o número de termos d, e a probabilidade do jth cara aparecer é dado por q_{jr} para o dado pertencente ao r^a classe. Portanto, se o dado for lançado eu 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 $\overline{Z}$ É dado por $(z_1 \ldots z_d)$, então a probabilidade gerativa do euo documento é dado pela seguinte distribuição multinomial:

$$P(\overline{Z} \mid C_r) = \frac{(\sum_{j=1}^{d} z_j)!}{z_1! z_2! \ldots z_d!} \prod_{j=1}^{d} (q_{jr})^{z_j} \propto \prod_{j=1}^{d} (q_{jr})^{z_j} \tag{5,18}$$

A constante de proporcionalidade se mantém C para $\overline{Z}$ fixo e classe variável, porque depende apenas de $\overline{Z}$ e é independente da classe r.

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 probabilidade de que a instância de teste $\overline{Z}$ pertence à classe C_r:

$$P(C_r \mid \overline{Z}) \propto P(C_r) \cdot P(\overline{Z} \mid C_r) \propto \alpha_r \prod_{j=1}^{d} (q_{jr})^{z_j} \tag{5,19}$$

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 Z.

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

$$q_{jr} = \frac{v(j, r)}{\sum_{j=1}^{d} v(j, r)} \tag{5,20}$$

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 $\gamma > 0$ para o numerador e $\gamma \cdot d$ para o denominador. Isso resulta na seguinte estimativa:

$$q_{jr} = \frac{v(j, r) + \gamma}{\sum_{j=1}^{d} v(j, r) + \gamma \cdot d} \tag{5,21}$$

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. Vários insights são fornecidos em [140] sobre por que a suposição ingênua funciona tão bem na prática.

Uma questão natural surge quanto a quando é preferível usar o modelo de Bernoulli ou o modelo multinomial. 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.

5.3.4 Classificação de resultados com Na¨ıve Bayes

O problema de predição de classificação nem sempre é colocado em termos de seleção da classe de uma única instância de teste. Em muitos casos, um conjunto de instâncias de teste $\overline{Z_1 \ldots Z_n}$ é fornecido, e