Aprendizado Bayesiano

Edigleison F. Carvalho
edigleison.carvalho@inf.ufrgs.br

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Motivação

- Considere um robô. Para se comportar de forma inteligente, o robô deve ser capaz de representar suas crênças sobre proposições no mundo:

  - “minha estação de recarga fica na localização (x,y,z)"
  - “meu sonar está com defeito"
  - “aquele stormtrooper é hostil"

Queremos representar a intensidade dessas crenças numericamente e conhecer as regras matemáticas para manipular essas crenças.

Motivação

- Com técnicas de aprendizado Bayesiano podemos fazer inferências probabilísticas sobre crenças a partir das evidências observadas no mundo.
- Premissa básica:
  - Quantidades de interesse são governadas por distribuições de probabilidade
  - Decisões ótimas podem ser feitas através do “raciocínio” sobre essas probabilidades juntamente com os dados de treinamento observados
- Aprendizado Bayesiano é relevante por dois motivos:
  - Primeiro porque permite a manipulação explícita de probabilidades; estando entre as abordagens mais práticas para certos tipos de problemas de aprendizagem; o classificador Bayesiano é competitivo com árvores de decisão e redes neurais artificiais.
  - Segundo porque fornecem uma perspectiva útil para entender métodos de aprendizado que não manipulam probabilidades explicitamente (ex.: justificar as funções de erro em redes neurais)

Características do Aprendizado Bayesiano

- Cada exemplo de treinamento observado pode aumentar ou diminuir a probabilidade estimada de que uma dada hipótese (crença) sobre o mundo seja correta
- Conhecimento a priori pode ser combinado com dados observados para determinar a probabilidade final de uma hipótese
- Métodos bayesianos trabalham com hipóteses que fazem previsões probabilísticas
- Novas instâncias podem ser classificadas combinando predições de múltiplas hipóteses, ponderadas pelas suas probabilidades

- Dificuldades práticas:
  - É requerido o conhecimento a priori de muitas probabilidades, que freqüentemente são estimadas a partir de conhecimento prévio, ou de dados previamente disponíveis, ou por suposições sobre a forma das distribuições
  - Requer custo computacional significante

Revisão - Probabilidade Condicional

- A inferência probabilística usa a informação disponível sobre os valores de algumas variáveis para obter a probabilidade para valores de outras variáveis.
- A probabilidade condicional *P*(*B*|*A*) é a probabilidade de *B* ocorrer dado que *A* tenha ocorrido. Ela é definida a partir da probabilidade conjunta *P*(*A*,*B*):

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

- *P*(*B*|*A*) é chamada de probabilidade *a posteriori* de *B*, no sentido de que a probabilidade de ocorrência de *B* é modificada *depois* que se constata a ocorrência de *A*, em relação ao que se sabia *a priori*, *P*(*B*).
- Esta probabilidade pode ser interpretada também como um *fator de confiança* que se pode inferir a partir dos dados na relação causal correspondente à regra: $A \rightarrow B$.
- *P*(*A*) e *P*(*B*) são as *probabilidades marginais* de *A* e *B*, respectivamente. Elas são também chamadas de probabilidades *a priori* destes valores de variáveis.
- A dificuldade do cálculo de *P*(*B*|*A*) está na determinação das probabilidades conjuntas.

Revisão – Probabilidade Condicional - Exemplo

- Considere os dados da tabela ao lado.

- Qual a probabilidade do tempo está nublado dado que Joga=sim?

$$P(nublado|sim) = \frac{4}{9}$$

- Qual a probabilidade do tempo está ensolarado dado que joga=nao?

$$P(ensolarado|nao) = \frac{3}{5}$$

- Qual a probabilidade do tempo está ensolarado e a temperatura está amena, isto é, quanto é $P(ensolarado, amena)$?

Tempo	Temperatura	Umidade	Ventoso	Joga
ensolarado	quente	alta	falso	não
ensolarado	quente	alta	verdadeiro	não
nublado	quente	alta	falso	sim
chuvoso	amena	alta	falso	sim
chuvoso	fria	normal	falso	sim
chuvoso	fria	normal	verdadeiro	não
nublado	fria	normal	verdadeiro	sim
ensolarado	amena	alta	falso	não
ensolarado	fria	normal	falso	sim
chuvoso	amena	normal	falso	sim
ensolarado	amena	normal	verdadeiro	sim
nublado	amena	alta	verdadeiro	sim
nublado	quente	normal	falso	sim
chuvoso	amena	alta	verdadeiro	não

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

inf UFRGS

Modelagem probabilística

- Um modelo descreve os dados que podem ser observados em um sistema
- Se utilizarmos teoria da probabilidade para expressar todas as formas de incerteza e ruídos associados com nosso modelo...
- então, a probabilidade inversa (regra de Bayes) permite-nos inferir quantidades desconhecidas, adaptar nossos modelos, fazer predições e aprender a partir dos dados

Regra de Bayes

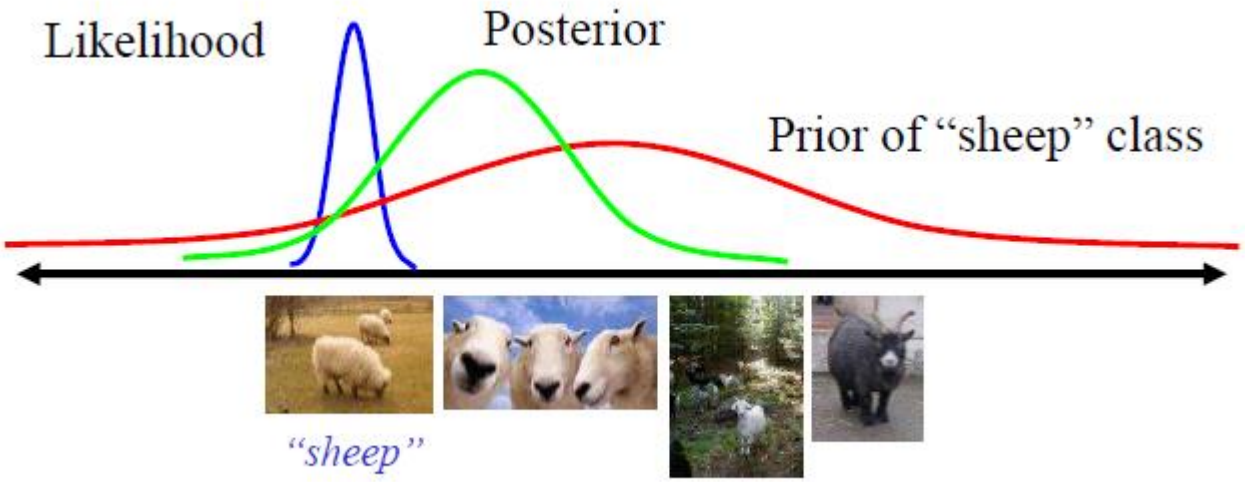
$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Thomas Bayes (1702–1761)

- $P(h)$: Probabilidade a priori da hipótese h
- $P(D)$: Probabilidade a priori dos dados de treinamento D
- $P(h|D)$: Probabilidade de h dado D ***(posterior prob.)***
- $P(D|h)$: Probabilidade de D dado h (***likelihood***)

- A regra (ou teorema) de Bayes nos diz como fazer inferência sobre hipóteses a partir dos dados.
- O teorema de Bayes mostra uma maneira prática de se calcular a probabilidade de uma hipótese/evento em particular, a partir de um conjunto de observações, sem a necessidade de se conhecer as probabilidades conjuntas.
- Aprendizado e predição podem ser vistos como formas de inferência.

Aprendizado e Inferência Bayesiana



MAP

- Normalmente queremos escolher a hipótese mais provável dado o conjunto de treinamento. Isto corresponde a hipótese com máxima probabilidade a posteriori (*maximum a posteriori - MAP*):

$$h_{MAP} = arg \max_{h \in H} P(h|D)$$

$$h_{MAP} = arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$h_{MAP} = arg \max_{h \in H} P(D|h)P(h)$$

Não é necessário calcular *P*(*D*), pois ela é a mesma para todas as hipóteses

Classificação por MAP

- Em tarefas de classificação, podemos representar as hipóteses do nosso modelo como sendo as diferentes classes $C_1 \dots C_k$. Pela regra de Bayes, a probabilidade a posteriori da classe C_i dado um vetor de entrada x é:

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

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{\sum_{j=1}^{k} P\left(x\middle|C_j\right)P(C_j)}$$

- Podemos usar MAP para classificar o vetor x como pertencente a uma das classes:

$$C_{MAP} = arg\max P(x|C_i)P(C_i)$$

$$C_{MAP} = arg\max logP(x|C_i) + logP(C_i)$$

Classificação por ML

- As vezes é assumido que as hipóteses são equiprováveis a priori
- Neste caso, a equação anterior é simplificada para:

$$C_{ML} = arg\max P(x|C_i)$$

- $P(x|C_i)$ é normalmente chamada de verossimilhança (*likelihood*) de x dado C_i. Qualquer hipótese *C* que maximiza $P(x|C)$ é chamada de hipótese de máxima verossimilhança (*maximum likelihood - ML*)

Exemplo

- Considere um problema de diagnóstico médico no qual existem duas possíveis hipóteses: (1) o paciente tem um certo tipo de câncer e (2) o paciente não tem câncer
- Os dados de um laboratório particular apontam que:

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$
$$P(+|cancer) = 0.98 \qquad P(-|cancer) = 0.02$$
$$P(+|\neg cancer) = 0.03 \qquad P(-|\neg cancer) = 0.97$$

- Suponha que o exame de um novo paciente deu positivo, devemos diagnosticar que o paciente tem câncer ou não?

$$P(+|cancer)P(cancer) = 0.98 \times 0.008 = 0.0078$$
$$P(+|\neg cancer)P(\neg cancer) = 0.03 \times 0.992 = 0.0298$$
$$h_{MAP} = \neg cancer$$

Exemplo

- As probabilidades a posteriori para o exemplo anterior podem ser calculadas normalizando os valores calculados para somar 1:

$$P(cancer|+) = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$

$$P(\neg cancer|+) = \frac{0.0298}{0.0078 + 0.0298} = 0.79$$

- Os resultados da inferência Bayesiana dependem fortemente das probabilidades a priori, as quais devem estar disponíveis para aplicação direta do método

Classificador ótimo Bayesiano

- A classificação mais provável é dada pelo classificador ótimo bayesiano que pondera as previsões de todas as hipóteses pelas respectivas probabilidades a posteriori.
- A probabilidade $P(v_j|D)$ que a classificação correta para x seja v_j é calculada por:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

- A classificação ótima para x é o valor de v_j para o qual $P(v_j|D)$ é máximo:

$$v(x) = arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Classificador ótimo Bayesiano

- No exemplo: $H = \{h_1, h_2, h_3\}$; $V = \{v_1 = +, v_2 = -\}$

$$P(h_1|x) = 0.4 \qquad P(+|h_1) = 1 \qquad P(-|h_1) = 0$$

$$P(h_2|x) = 0.3 \qquad P(+|h_2) = 0 \qquad P(-|h_2) = 1$$

$$P(h_3|x) = 0.3 \qquad P(+|h_3) = 0 \qquad P(-|h_3) = 1$$

$$P(v_1|D) = \sum_{h_i \in H} P(v_1|h_i)P(h_i|D) = 0.4$$

$$P(v_2|D) = \sum_{h_i \in H} P(v_2|h_i)P(h_i|D) = 0.3 + 0.3 = 0.6$$

$$v(x) = arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = v_2 = -$$

Estimação e Inferência de Parâmetros

- Os métodos de aprendizado Bayesiano se dividem em duas categorias que dizem respeito sobre a quantidade de parâmetros utilizados:
  - Abordagens **paramétricas:** tem um número fixo (e finito) de parâmetros, independente do tamanho do conjunto de treinamento. Dado Θ, as predições são independentes dos dados *D:*

$$p(x, \Theta|D) = p(x|\Theta)p(\Theta|D)$$

    Os parâmetros são um resumo dos dados. Podemos também chamar de aprendizado baseado em modelo (*model-based learning*). Ex.: Mistura de Gaussianas
  - Abordagens **não-paramétricas**: permite que o número de parâmetros cresçam com o tamanho do conjunto de treinamento, ou alternativamente podemos pensar que as predições são dependentes dos dados, e possivelmente de um conjunto pequeno de parâmetros α

$$p(x|D, \alpha)$$

    Podemos chamar isso de aprendizado baseado em memória (*memory-based learning*). Ex.: estimadores de kernel, Processo de Dirichlet Hierárquico

Estimação e Inferência de Parâmetros

- Na abordagem *paramétrica*, assume-se que a amostra é obtida de uma distribuição que obedece a um modelo conhecido, p. ex. gaussiano.

- Neste caso, a especificação do modelo se dá pela *estimação* dos valores de um número pequeno de parâmetros, a *estatística suficiente* da distribuição (p. ex., média e variância).

- Uma vez que estes parâmetros sejam estimados a partir de uma amostra, toda a distribuição se torna conhecida.

- A *estimação por máxima verossimilhança* é o método fundamental para estimar os parâmetros de uma distribuição

Estimação por Máxima Verossimilhança (MLE)

- Digamos que temos uma amostra independente e identicamente distribuída (iid) $X = [x_1, \ldots, x_n]$
- Assumimos que $x_i \sim p(x|\Theta)$ sejam instâncias retiradas de uma família de *densidades de probabilidade*, $p(x|\Theta)$, definida pelos parâmetros Θ.
- Desejamos encontrar o Θ que torne a amostragem de x_i por $p(x|\Theta)$ o mais provável possível.
- Como os x_i são independentes, a probabilidade do parâmetro Θ dada a amostra *X*, a chamada *verossimilhança* (*likelihood*) de Θ, é o produto das verossimilhanças de cada instância:

$$l(\Theta|X) \equiv p(X|\Theta) = \prod_{i=1}^{n} p(x_i|\Theta)$$

Estimação por Máxima Verossimilhança (MLE)

- Na *estimação por máxima verossimilhança*, desejamos encontrar Θ que maximiza a verossimilhança dos dados $l(\Theta|X)$, que é equivalente a maximizar o seu logaritmo:

$$\log l(\Theta|X) = \sum_{i=1}^{n} p(x_i|\Theta)$$

- Se *X* seguir uma distribuição normal (gaussiana) com média $E[X] = \mu$ e *variância* σ^2, a sua função de densidade de probabilidade é:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$$

Estimação por Máxima Verossimilhança (MLE)

- Neste caso, desejamos encontrar Θ que maximiza:

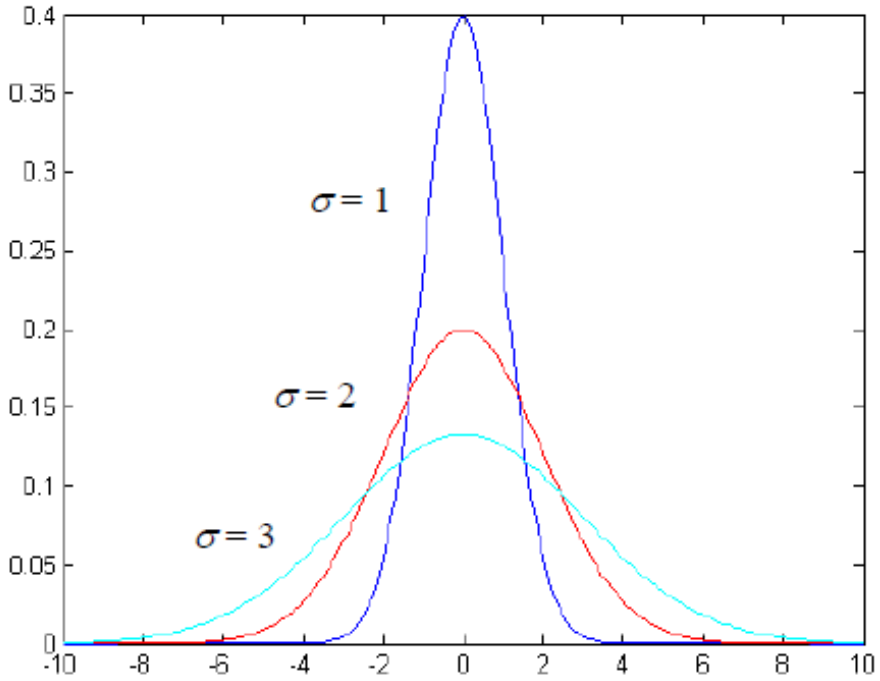
$$\log l(\Theta|X) = -\frac{N}{2}\log(2\pi) - N.\log\sigma - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

- As estimativas da média μ e variância σ^2, por MLE, são obtidas tomando as derivadas parciais de $\log l(\Theta|X)$ em relação aos parâmetros, e igualando a 0, resultando em:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Exemplos de distribuições normais



$\mu = 0$

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Classificador Naive Bayes

- O classificador Naive Bayes (NB), também chamado de Classificador Bayesiano Ingênuo é um dos mais simples, porém pode ser eficiente ou razoável para certos tipos de dados.
- O classificador Naive Bayes assume que os *D* atributos/*features* de uma amostra *x* são condicionalmente independentes dado um rótulo de classe *c*
- Isso nos permite reescrever a verossimilhança (*likelihood*) de um vetor *x* como:

$$p(x|c,\Theta) = \prod_{j=1}^{D} p(x_j|c,\Theta_{jc})$$

- Normalmente o classificador Naïve Bayes funciona bem mesmo quando sua premissa não é verdade.
- Uma razão para isto é que o modelo é muito simples (ele tem somente *O*(*CD*) parâmetros, para *C* classes e *D* atributos), e, portanto, relativamente imune a *overfitting*.

Classificador Naive Bayes

- A forma da verossimilhança depende da distribuição de probabilidade que assumimos que os dados seguem. Por exemplo, para atributos contínuos ou reais, podemos utilizar a distribuição Gaussiana:

$$p(x|c,\Theta) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_{jc},\sigma_{jc}^2)$$

onde μ_{jc} é a média do atributo *j* nos objetos da classe *c,* e σ_{jc}^2 é sua variância.

- Para atributos binários, $x_j \in \{0,1\}$, podemos utilizar a distribuição de Bernoulli.
- E para atributos categóricos ou discretos com *K* valores, , $x_j \in \{1,\ldots,K\}$, podemos utilizar a distribuição multinomial.
- Podemos utilizar diferentes premissas sobre a distribuição dos dados para tratar certos tipos de atributos e também é fácil misturar diferentes tipos de atributos

Exemplo

- Dado a nova observação: <ensolarado,fria,alta,verdadeiro>, qual a classificação por NB?

$$P(sim).P(ensolarado|sim).P(fria|sim).P(alta|sim).P(verdadeiro|sim)$$

= 9/14.2/9.3/9.3/9.3/9 = 0.0053

$$P(nao).P(ensolarado|nao).P(fria|nao).P(alta|nao).P(verdadeiro|nao)$$

= 5/14.3/5.1/5.4/5.3/5 = 0.0206

MAP: não joga!

Normalizando:

$$\frac{0.0206}{0.0206 + 0.0053} = 0.795$$

Tempo	Temperatura	Umidade	Ventoso	Joga
ensolarado	quente	alta	falso	não
ensolarado	quente	alta	verdadeiro	não
nublado	quente	alta	falso	sim
chuvoso	amena	alta	falso	sim
chuvoso	fria	normal	falso	sim
chuvoso	fria	normal	verdadeiro	não
nublado	fria	normal	verdadeiro	sim
ensolarado	amena	alta	falso	não
ensolarado	fria	normal	falso	sim
chuvoso	amena	normal	falso	sim
ensolarado	amena	normal	verdadeiro	sim
nublado	amena	alta	verdadeiro	sim
nublado	quente	normal	falso	sim
chuvoso	amena	alta	verdadeiro	não

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

inf UFRGS

Estimando probabilidades

- Se um valor de atributo nunca ocorrer para uma classe (Tempo nublado para “não”)?
  - A probabilidade será zero! *P*(*nublado* | *não*) = 0
  - A probabilidade a posteriori será zero, independentemente dos outros valores!
- Solução: *Estimador de Laplace* -> somar 1 à contagem de todas as combinações de classe e valor de atributo.
- Resultado: as probabilidades nunca serão zero!
- Equivale a acrescentar *m* = |*E*| amostras de cada classe aos dados de treinamento $P(x_i|c) = (n_c + 1)/(n + m)$

  n : número de instâncias da classe *c*

  n_c: número de exemplos da classe *c* com o valor x_i

  m : |*E*|, número de valores do atributo *E*

Exemplo

- Dado a nova observação: <nublado,fria,alta,verdadeiro>, qual a classificação por NB?

$$P(sim).P(nublado|sim).P(fria|sim).P(alta|sim).P(verdadeiro|sim)$$

= 9/14.5/12.4/12.4/11.4/11 = 0.0118

$$P(nao).P(nublado|nao).P(fria|nao).P(alta|nao).P(verdadeiro|nao)$$

= 5/14.1/8.2/8.5/7.4/7 = 0.0045

MAP: joga!

Normalizando:

$$\frac{0.0118}{0.0118 + 0.0045} = 0.724$$

Tempo	Temperatura	Umidade	Ventoso	Joga
ensolarado	quente	alta	falso	não
ensolarado	quente	alta	verdadeiro	não
nublado	quente	alta	falso	sim
chuvoso	amena	alta	falso	sim
chuvoso	fria	normal	falso	sim
chuvoso	fria	normal	verdadeiro	não
nublado	fria	normal	verdadeiro	sim
ensolarado	amena	alta	falso	não
ensolarado	fria	normal	falso	sim
chuvoso	amena	normal	falso	sim
ensolarado	amena	normal	verdadeiro	sim
nublado	amena	alta	verdadeiro	sim
nublado	quente	normal	falso	sim
chuvoso	amena	alta	verdadeiro	não

Tratando valores faltantes

- No treinamento: o exemplo não é incluído na contagem de frequências para a combinação de classe-valor de atributo
- Na classificação: atributo será omitido do cálculo
- Ex.: dado observado: *x* = <?, fria, alta, verdadeiro>, joga?

$$P(sim).P(fria|sim).P(alta|sim).P(verdadeiro|sim)$$
$$= \frac{9}{14}.\frac{3}{9}.\frac{3}{9}.\frac{3}{9} = 0{,}0238$$

$$P(nao).P(fria|nao).P(alta|nao).P(verdadeiro|nao)$$
$$= \frac{5}{14}.\frac{1}{5}.\frac{4}{5}.\frac{3}{5} = 0{,}0343$$

Normalizando:

$$P(sim|x) = \frac{0{,}0238}{0{,}0238 + 0{,}0343} = 0{,}41$$

$$P(nao|x) = \frac{0{,}0343}{0{,}0238 + 0{,}0343} = 0{,}59$$

Classificação de documentos

- Objetivo: aprender quais documentos são de interesse (duas classes: “interesse” , “não-interesse”).
- O classificador Naive Bayes é um dos métodos mais efetivos nesta tarefa.
- As *n* palavras $a_1, \ldots, a_n$ num documento são realizações de palavras de um vocabulário *V* de tamanho *K*, com $V = \{w_1, \ldots, w_K\}$ O vocabulário *V* é aprendido extraindo as distintas palavras de todos os documentos do conjunto de treinamento.
- A probabilidade a priori das classes é definida por:

$$P(+) = \frac{< \#\ de\ documentos\ da\ classe\ + >}{número\ total\ de\ documentos}$$

e

$$P(-) = \frac{< \#\ de\ documentos\ da\ classe\ - >}{número\ total\ de\ documentos}$$

Classificação de documentos

- A abordagem do Naive Bayes considera que as posições das palavras são independentes dado o rótulo de classe. Neste caso, a probabilidade condicional dado a classe é obtido por:

$$P(doc|C) = \prod_{i=1}^{n} P(a_i = w_k|C)$$

onde $P(a_i = w_k|C)$ é a probabilidade que a palavra na posição *i* seja w_k, dado *C* (+ ou -). $P(a_i = w_k|C)$ pode ser calculada como a frequência com que a palavra w_k aparece entre os documentos da classe *C.*

- Como uma palavra contida no vocabulário pode não aparecer num dado documento, então torna-se necessário o uso de alguma prior sobre esse atributo. Para tanto, pode-se utilizar o estimador de Laplace:

$$P(a_i = w_k|C) = \frac{n_{ij} + 1}{n_j + K}$$

onde n_j é o número total de palavras nos documentos da classe *C* e n_{ij} é o número de ocorrências da palavra w_k nos documentos da classe *C*

Classificação de documentos

- Pseudocódigo do classificador NB para textos (aprendizado):

Input: Docs, Classes

1. Colete todas as palavras e outros *tokens* que ocorram nos documentos Docs e defina o vocabulário *V* como todas as distintas palavras e outros *tokens.*
2. Calcule as probabilidades a priori para cada classe *C* e as probabilidades condicionais $P(w_k|C)$ para cada palavra w_k de *V.*

- Para cada classe *C* faça:

$docs_j \leftarrow subconjunto\ de\ Docs\ pertencentes\ a\ classe\ C$

$$P(C) = \frac{<\#\ de\ documentos\ da\ classe\ C>}{número\ total\ de\ documentos}$$

$Text_j = \text{único}\ documento\ criado\ pela\ concatenação\ de\ todos\ os\ docs_j$

$n \leftarrow número\ total\ de\ palavras\ em\ Text_j$

para cada palavra w_k no vocabulário

$n_k \leftarrow número\ de\ vezes\ que\ a\ palavra\ w_k\ ocorre\ em\ Text_j$

$$P(w_k|C) \leftarrow \frac{n_k+1}{n+K}$$

Classificação de documentos

- Classificação por MAP:

Input: novo documento

Return: $arg \max_C P(C).\prod_{i=1}^{n} P(a_i|C)$

20 NewsGroups

- Dado 1000 documentos de treinamento de cada grupo, aprenda a classificar novos documentos como pertencentes a um dos grupos

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Acurácia do NB: **89**%

Artigo do rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!ogicse!uwm.edu
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinion)...
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
```

NB para Dígitos

- Considere uma versão simplificada do problema de classificação de dígitos em que os pixels assumem valores binários {0,1} indicando se o pixel está preto/*on* ou não.
- Uma feature F_{ij} para cada posição (*i*,*j*) na grade.
- Cada imagem é mapeada para um vetor:

$$\rightarrow \langle F_{0,0} = 0 \;\; F_{0,1} = 0 \;\; F_{0,2} = 1 \;\; F_{0,3} = 1 \;\; F_{0,4} = 0 \;\; \ldots F_{15,15} = 0 \rangle$$

- Modelo NB:

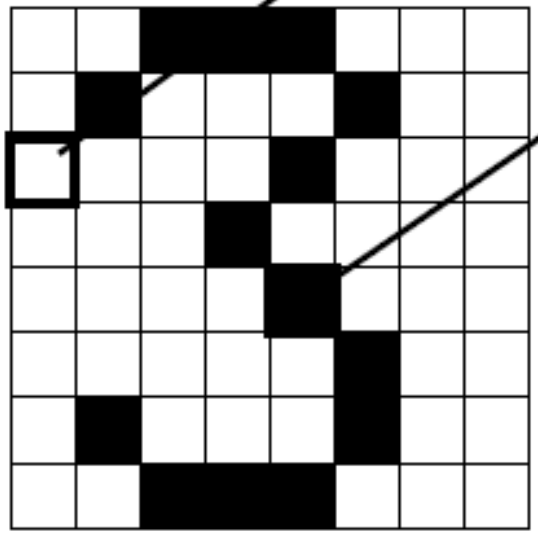
$$P(C|F_{0,0}, \ldots, F_{15,15}) \propto P(C) \prod_{i,j} P(F_{i,j}|C)$$

- As features são independentes dado a classe?

NB para Dígitos

P(C)

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



$P(F_{3,1} = on|Y)$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$P(F_{5,5} = on|Y)$

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

NB Overfitting

$P(\text{features}, C = 2)$

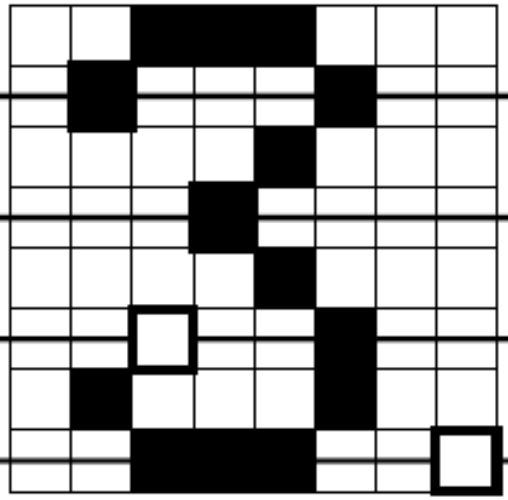
$P(C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.8$

$P(\text{on}|C = 2) = 0.1$

$P(\text{off}|C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.01$



$P(\text{features}, C = 3)$

$P(C = 3) = 0.1$

$P(\text{on}|C = 3) = 0.8$

$P(\text{on}|C = 3) = 0.9$

$P(\text{off}|C = 3) = 0.7$

$P(\text{on}|C = 3) = 0.0$

2 vence!

NB gaussiano

- Se a entrada for contínua se deve utilizar o NB com distribuição gaussiana.

Exemplos de imagens de caracteres com valores reais nos pixels

$$p(x|c,\Theta) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_{jc},\sigma_{jc}^2)$$

- Ou em logaritmo para evitar erros numéricos com valores muito pequenos de probabilidades:

$$p(x|c,\Theta) = \sum_{j=1}^{D} \log \mathcal{N}(x_j|\mu_{jc},\sigma_{jc}^2)$$

Discussão da classificação do NB

- NB funciona muito bem, mesmo se a suposição de independência for violada.
- Por quê? Porque a classificação não requer estimativas precisas de probabilidades contanto que a máxima probabilidade for atribuída à classe correta
- Entretanto: adicionar muitos atributos redundantes causa problemas (atributos idênticos)
- Também: muitos atributos numéricos não têm distribuição normal (-> estimadores de densidade por kernel)