بسم الله الرحمن الرحيم

الدكتور هيتكلم انهارده علي naïve Bayes و هنتكلم علي الفرق ما بين ال Generative models ... حاجه زي naïve bayes ... و ال Discriminative models .. حاجه زي ال logistic regression .. في سؤال .. ليه بنستخدم Naïve bayes .. الاجابه ان ليه ميزة وهو ان ال training بيبقا سريع جدا ... اللي تعملو بس هو انك تحسب ال parameters in closed form عن طريق انك تعد بس .. في بقيت الطرق انت حتاج وقت كبير عشان تعمل ال training و بتفضل تروح وتعيد في ال epochs وكدا .. طيب يلا بينا

Naïve bayes model ده بيوصف Generative model .. بمعني .. كلمة generative هنا معناها هو وصف لإزاي احنا نقدر ن generate the data we are working wiith .. قبل كدا انت مكنش عندك الحوار ده ... انت كان آخرك في ال discriminative model هو ان الإكسات تبقا عندك .. وبعدين ال y بقا هنا كان ممكن تقول ال y ديه generative . جات منين يعني بس ال y بتبقا coniditioned on x ..... هنا انت عندك في ال generative ازاي ال x and y بيبقو generated .. ففي الصوره .. انت بت flip the coin .. خد القيمه بتاعته اللي هي هتبقا ال y .. وبعدين رووح اعمل flipping to the left/right coins .. علي حسب قيمة الواي لو 1 روح اقلب اللي لونهم أزرق .. لو صفر روح اقلب اللي لونهم احمر ... الفكره ان في assumption بتاع Naïve Bayes ان الفيتشرز بتبقا independend ... اللي عاوزين نعملو ناو اننا نعمل ملخص كدا سريع لل clsed form MLE applied to NVBayes model ....

generation of y values

g

generation of x values

(2) Log-likelihood:

$$\ell(\phi,\Theta) = \log \prod_{i=1}^{N} p(x^{(i)}, y^{(i)} \mid \phi, \Theta)$$

$$= \sum_{i=1}^{N} \left[\log p(y^{(i)} \mid \phi) + \sum_{m=1}^{M} \log p(x_m^{(i)} \mid y^{(i)}, \Theta) \right]$$

$$= N_{y=H} \log \phi + N_{y=T} \log(1-\phi)$$
$$+ \sum_{m=1}^{M} N_{x_m=1,y=H} \log(\theta_{H,m}) + N_{x_m=0,y=H} \log(1-\theta_{H,m})$$
$$+ \sum_{m=1}^{M} N_{x_m=1,y=T} \log(\theta_{T,m}) + N_{x_m=0,y=T} \log(1-\theta_{T,m})$$

$N_{y=H} = \#(y^{(i)}=H)$

$N_{y=T} = \#(y^{(i)}=T)$

$N_{x_m=1,y=H}$ = # times that $x_m^{(i)}=1$ and $y^{(i)}=H$

M features

$y^{(i)}=H$

$y^{(i)}=T$

$x_m^{(i)}=1$

$x_m^{(i)}=0$

Case a) ϕ b) Θ elements

(3a) Take partial derivatives w.r.t. ϕ

$$\frac{\partial \ell(\phi,\Theta)}{\partial \phi} = \frac{N_{y=1}}{\phi} - \frac{N_{y=0}}{1-\phi}$$

(4a) Set to zero and solve for ϕ

$$\phi^{MLE} = \frac{N_{y=1}}{N_{y=1}+N_{y=0}} = \frac{N_{y=1}}{N}$$

big gold coin

No. of times it was heads

(3b) Take partial derivative wrt $\Theta_{H,m}$ (case for $y=H$ and feature m)

$$\frac{\partial \ell(\phi,\Theta)}{\partial \Theta_{H,m}} = \frac{N_{x_m=1,y=H}}{\Theta_{H,m}} - \frac{N_{x_m=0,y=H}}{1-\Theta_{H,m}}$$

(4b) Solve for $\Theta_{H,m}$

$$\Theta_{H,m}^{MLE} = \frac{N_{x_m=1,y=H}}{N_{x_m=1,y=H} + N_{x_m=0,y=H}} = \frac{N_{x_m=1,y=H}}{N_{y=H}}$$

For red coins, H and blue coins, T

weights of Coins by Counting

طيب دلوقت احنا معانا ال Naïve bayes max likelihood estimate .. احنا هنفكر شويه في تفاصيل الموديل بتاع Bernoulli Naïve bayes ..

Model 1: Bernoulli Naïve Bayes

X is binary feature vector

y is binary label

Data: Binary feature vectors, Binary labels

$\mathbf{x} \in \{0,1\}^M$ $\quad$ $y \in \{0,1\}$

Generative Story:

$y \sim \text{Bernoulli}(\phi)$

$x_1 \sim \text{Bernoulli}(\theta_{y,1})$

$x_2 \sim \text{Bernoulli}(\theta_{y,2})$

$\vdots$

$x_M \sim \text{Bernoulli}(\theta_{y,M})$

Model:

$$\begin{aligned} p_{\phi,\boldsymbol{\theta}}(\boldsymbol{x}, y) &= p_{\phi,\boldsymbol{\theta}}(x_1, \ldots, x_M, y) \\ &= p_\phi(y) \prod_{m=1}^{M} p_{\boldsymbol{\theta}}(x_m|y) \\ &= \Big[(\phi)^y (1-\phi)^{(1-y)} \\ &\quad \prod_{m=1}^{M} (\theta_{y,m})^{x_m} (1-\theta_{y,m})^{(1-x_m)}\Big] \end{aligned}$$

Model 1: Bernoulli Naïve Bayes

Maximum Likelihood Estimation ← عشان نفصل

Training: Find the **class-conditional** MLE parameters

Count Variables:

$$N_{y=1} = \sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 1)$$

$$N_{y=0} = \sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 0)$$

$$N_{y=0,x_m=1} = \sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 0 \wedge x_m^{(i)} = 1)$$

$$\ldots$$

Maximum Likelihood Estimators: → algebraic Representation of How we can define ϕ & θ

$$\phi = \frac{N_{y=1}}{N}$$

$$\theta_{0,m} = \frac{N_{y=0,x_m=1}}{N_{y=0}}$$

$$\theta_{1,m} = \frac{N_{y=1,x_m=1}}{N_{y=1}}$$

$$\forall m \in \{1, \ldots, M\}$$

17

A Shortcoming of MLE

For Naïve Bayes, suppose we **never** observe the word "unicorn" in a real news article.

In this case, what is the MLE of the following quantity?

$\boxed{p(x_{unicorn} = 1 \mid y = real) = 0}$

y=Real & $x_{unicorn} = 1$

Recall: MLE $\theta_{k,0} = \dfrac{\sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 0 \wedge x_k^{(i)} = 1)}{\sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 0)}$ ← tot # of y, Real

Now suppose we observe the word "unicorn" at test time. What is the posterior probability that the article was a real article?

test doc. contains unicorn

$$p(y = real \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y = real)\, p(y = real)}{p(\mathbf{x})}$$

$p(y = real|\mathbf{x}) = 0$, $p(\mathbf{x}|y=real) = \prod_{m=1}^{M} p(x_m|y=real)$, $p(y=real) = ?$, $p(\mathbf{x}) = ?$

هو ال posterior = 0 عشان احنا ملاحظناش الكلمه فى اى حتة.

Posterior Prob → y = Real|x

23

الدكتور في السلايد ديه بيقول بعد ما لقي ان ال posterior هتبقا بصفر عشان احنا بس ملاحظناش الكلمه اللي اسمها unicorn قبل كدا في أي حاجه Real .. فال posterior بقا بصفر .. فال document مبقتش real .. في الحاله ديه في حاجه مش منطقيه .. انك تقول ان document isn't real عشان مجرد ان في كلمه نادراً ما كانت موجوده في كل ال articles اللي شفناها في الترينينج تايم ... يبقا كان احسن ان ال p(x|y=real) تبقا قيمتها صغيره مش تبقي بصفر علي طول كدا .. ايه الي هيحصل لو كل الترمات الي في ال p(x|y=real) اللي هي عباره عن ضرب من واحد لحد M لل p(x_m|y=real) .. ايه اللي هيحصل لو ان كل الترمات ديه كانت ليهم high probability ماعدا الترم الي اسمو unicorn ..دلوقت معناه ان عندك مشكله .. المشكله ديه بتبقا ان ال article للي انت ماسكها ف ايدك ديه ... كان شكلها فعلاً خلاص a real news article ... ماعدا لكلمة unicorn ... طيب اللي هنعملو دلوقت هو ازاي هنحل المشكله ديه .. ب ال MAP estimation .. وديه ال recipe بتاعتها:

Recipe for Closed-form MAP Estimation

1. Assume data was generated i.i.d. from some model (i.e. write the generative story) sampled
   $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and then for all i: $x^{(i)} \sim p(x|\boldsymbol{\theta})$
2. Write log-likelihood Prior
   $\ell_{MAP}(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log p(x^{(1)}|\boldsymbol{\theta}) + \ldots + \log p(x^{(N)}|\boldsymbol{\theta})$
3. Compute partial derivatives (i.e. gradient)
   $\partial \ell_{MAP}(\boldsymbol{\theta})/\partial \theta_1 = \ldots$
   $\partial \ell_{MAP}(\boldsymbol{\theta})/\partial \theta_2 = \ldots$
   $\ldots$
   $\partial \ell_{MAP}(\boldsymbol{\theta})/\partial \theta_M = \ldots$
4. Set derivatives to zero and solve for $\boldsymbol{\theta}$
   $\partial \ell_{MAP}(\boldsymbol{\theta})/\partial \theta_m = 0$ for all $m \in \{1, \ldots, M\}$
   $\boldsymbol{\theta}^{MAP}$ = solution to system of M equations and M variables
5. Compute the second derivative and check that $\ell(\boldsymbol{\theta})$ is concave down at $\boldsymbol{\theta}^{MAP}$

بدل ما بنشتغل باللوج لايكلي هوود بتاع الداتا .. بنشتغل بال L_MAP اللي بيضيف الترم بتاع ال log(prior) وبعديه بقيت ترمات ال log likelihood ..
وبعدين نحسب ال partial derivatives ...

$\phi \sim$ Uniform([0,1])

Model 1: Bernoulli Naïve Bayes

MAP Estimation (Beta Prior)

1. Generative Story:

The parameters are drawn once for the entire dataset.

for $m \in \{1, \ldots, M\}$:
  for $y \in \{0, 1\}$:
    $\theta_{y,m} \sim$ Beta(α, β)
for $i \in \{1, \ldots, N\}$:
  $y^{(i)} \sim$ Bernoulli(ϕ)
  for $m \in \{1, \ldots, M\}$:
    $x_m^{(i)} \sim$ Bernoulli$(\theta_{y^{(i)},m})$

$$N_{y=1} = \sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 1)$$

$$N_{y=0} = \sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 0)$$

$$N_{y=0,x_m=1} = \sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 0 \wedge x_m^{(i)} = 1)$$

$\ldots$

2. Likelihood:

$$\ell_{MAP}(\phi, \boldsymbol{\theta})$$

$$= \log [\underbrace{p(\phi, \boldsymbol{\theta}|\alpha, \beta)}_{\text{prior}} \underbrace{p(\mathcal{D}|\phi, \boldsymbol{\theta})}_{\text{like.}}]$$

$$= \log \left[\left(p(\phi|\alpha, \beta) \prod_{m=1}^{M} p(\theta_{0,m}|\alpha, \beta) \right) \left(\prod_{i=1}^{N} p(\mathbf{x}^{(i)}, y^{(i)}|\phi, \boldsymbol{\theta}) \right) \right]$$

$p(\theta_{1,m}|\alpha, \beta)$

3. MAP Estimators: $(\phi^{MAP}, \boldsymbol{\theta}^{MAP}) = \underset{\phi, \boldsymbol{\theta}}{\text{argmax}}\, \ell_{MAP}(\phi, \boldsymbol{\theta})$

Take derivatives, set to zero and solve...

$\alpha = 2$
$\beta = 2$

$$\phi = \frac{N_{y=1}}{N}$$

new Term different from MLE

$$\theta_{0,m} = \frac{(\alpha - 1) + N_{y=0,x_m=1}}{(\alpha - 1) + (\beta - 1) + N_{y=0}}$$

$$\theta_{1,m} = \frac{\overset{2-1=1}{(\alpha - 1)} + N_{y=1,x_m=1}}{(\alpha - 1) + (\beta - 1) + N_{y=1}}$$

$$\forall m \in \{1, \ldots, M\}$$

@ $\alpha, \beta = 2$
كدا انت البسط
كحزل كهروما
هيبقى بعض

كدا انت لما هتقابل كلمة unicron تاني .. هيبقا ليها low probability بس هتبقا overwhelmed ببقيت الكلمات اللي فعلاً بتخلي ال Doc اللي معاك دلوقت .. real news article .. وهو ده الهدف من ان احنا ليه بنعوز prior distribution of the parameters .. عشان ال prior بتقلك

"It's extremely unlikely that one of the parameters should be an exact zero"

Model 2: Multinomial Naïve Bayes

Support: Option 1: Integer vector (word IDs)

\# or Num of features

$$\mathbf{x} = [x_1, x_2, \dots, x_M] \text{ where } x_m \in \{1, \dots, K\} \text{ a word id.}$$

0, 1, 2, ..., K-1

each feature is a count

newname
10/20/2021 9:02:06 AM

ررقم من واحد ل K ... كدا الدوكيومنت ريبرزنتيشن مش هتبقا مجرد
انك تقول هل الكلمه ديه موجوده ولا لا .. انما هتقول
how many time does this word appear in that document

Generative Story:

for $i \in \{1, \dots, N\}$:

$y^{(i)} \sim \text{Bernoulli}(\phi)$

for $j \in \{1, \dots, M_i\}$:

$x_j^{(i)} \sim \text{Multinomial}(\theta_{y^{(i)}}, 1)$

newname
10/20/2021 9:04:09 AM

هنا لكل example هنسامبل ال y i من برنوللي
ديستربيوشن .. و هنسامبل كل فيتشر عندنا من
Multinomial distribution .. وده ان عندك k-sided
dice roll .. هتاخد dye و بعدين ت roll that dye و
هيطلع بقيمه و بالطريقه ديه انت بت generate value
لل x_j

Model:

$$p_{\phi,\boldsymbol{\theta}}(\boldsymbol{x}, y) = p_\phi(y) \prod_{k=1}^{K} p_{\boldsymbol{\theta}_k}(x_k | y)$$

$$= (\phi)^y (1-\phi)^{(1-y)} \prod_{j=1}^{M_i} \theta_{y, x_j}$$

الدكتور بيقول ان الموديل هو هو اللي كنا بنعملو قبل كدا . هنضرب
البروبابلتي بتاعت واي في (لكل فيتشر) اضرب احتمالية الفيتشر ديه
تحت شرط الواي
الفكره بس ان عندك
K sided dye for each of the possible outcomes of y

Bernolli ← لو عاوز Multi class نشيل
و يحط
Multinomial

$\boxed{\text{Gaussian N.B.}}$ (Multiclass)

Multinomial$(\vec{\phi}, 1)$

$y \sim$ ~~Bernoulli~~(ϕ) $= p(y|\phi)$ ← pmf

$x_1 \sim \text{Gaussian}(\mu_{y,1}, \sigma^2_{y,1}) = f(x_1 | \dots)$

$x_2 \sim \text{Gaussian}(\mu_{y,2}, \sigma^2_{y,2}) = f(x_2 | \dots)$

$\vdots$ $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \vdots$ pdf

$x_M \sim \text{Gaussian}(\mu_{y,M}, \sigma^2_{y,M}) = f(x_M | \dots)$

x_M → Mth features

in MAP likelihood
↓
Pmf * Pdfs

MAP Estimation and Regularization

MAP

$$\hat{\Theta}^{MAP} = \underset{\Theta}{\arg\max}\ \log p(D|\Theta) + \log p(\Theta)$$

$$= \arg\min \underbrace{-\log p(D|\Theta)}_{} \underbrace{-\log p(\Theta)}_{}$$

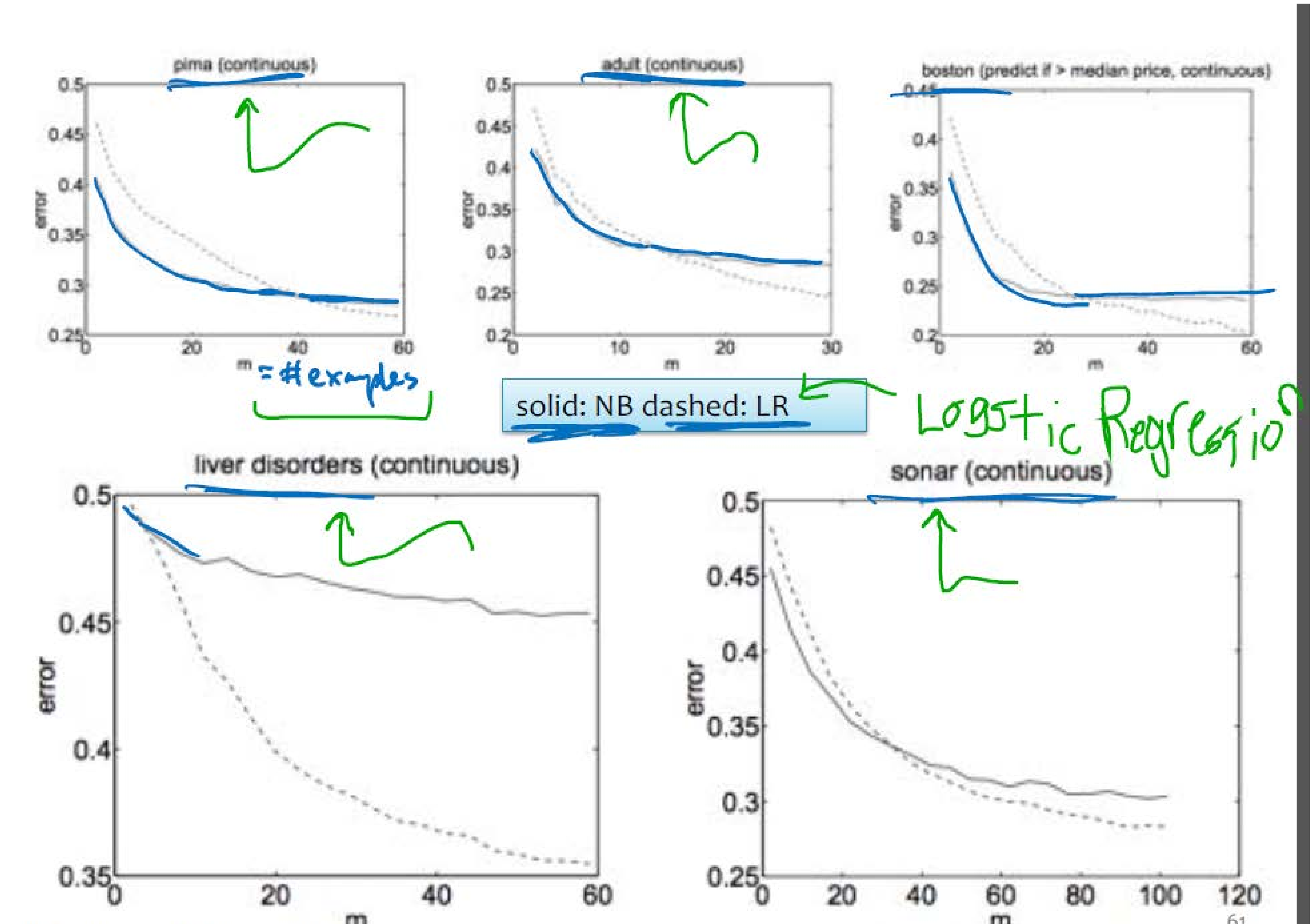
Regularization

$$\hat{\Theta} = \underset{\Theta}{\arg\min}\ \underbrace{J(\Theta|D)}_{\text{fit the data}} + \underbrace{r(\Theta)}_{\text{keep the model "simple"}}$$

→ Regularizer

show that ↓

L2 Reg. ≡ Gaussian Prior



Slide courtesy of William Cohen