بسم الله الرحمن الرحيم

الكورس هيبقا عنيف .. ربنا ييسر الأمور ... هومورك رقم 6 نزل يا بيييه ... خلص 5 انجز ... الدكتور عارف انك مش بتقرا ... اقرا يا بيييه .... طيب الدكتور بيقول الحاجه الي عاوزين نركز عليها هي ال Max likelihood estimation و ال Max A-posterior estimation .. عشان نعمل كدا .. احنا هنبدأ نفكر في ال likelihood function .. ال setting بتاعها ان عندك N data samples .. دول كلهم sampled from random variable X ... تعال نبص علي أول حاله .. الإكس ديسكريت ... الضرب هنا حصل عشان هنا انت حاطط شرط iid ... تاني حاله .. ال likelihood function بقت فانكشن في الدينستي .. في الحالتين .. ال likelihood بتقولنا how likely it is one sample to another .. فلما بنتكلم علي continuous random variable انت هنا مش بتضرب احتمالات فبعض .. عادي .. انت بس مهتم بال relative likelihood مابين ال different settings of the parameter ...

Likelihood Function

One R.V.

- Suppose we have N **samples** $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ from a **random variable** X
- The **likelihood** function: prob. mass function
  - Case 1: X is **discrete** with *pmf* $p(x|\theta)$
    $L(\theta) = p(x^{(1)}|\theta)\, p(x^{(2)}|\theta) \ldots p(x^{(N)}|\theta)$ indep. identical dist (iid)
  - Case 2: X is **continuous** with *pdf* $f(x|\theta)$ prob. density fn.
    $L(\theta) = f(x^{(1)}|\theta)\, f(x^{(2)}|\theta) \ldots f(x^{(N)}|\theta)$

In both cases (discrete / continuous), the **likelihood** tells us how likely one sample is relative to another

- The **log-likelihood** function:
  - Case 1: X is **discrete** with *pmf* $p(x|\theta)$
    $\ell(\theta) = \log p(x^{(1)}|\theta) + \ldots + \log p(x^{(N)}|\theta)$
  - Case 2: X is **continuous** with *pdf* $f(x|\theta)$
    $\ell(\theta) = \log f(x^{(1)}|\theta) + \ldots + \log f(x^{(N)}|\theta)$

طيب احنا هنتكلم علي ال log likelihood function .. خد اللوج بس .. الكلام ده كلو ل One random variable .. اللي احنا كنا بنتكلم فيه طول الترم اصلا هو لما بيكون عندنا 2 راندم فاريبلز .. فاحنا عندنا pairs x and y .. سحبناهم من random variables x and y .. اللي هنتكلم عنو انهارده هو ال joint likelihood function .. ده اللي هيطلعلنا لما نتكلم علي naïve Bayes ... عندنا حالتين برضو .. يا دسكريت يا كونتنوس ..

Likelihood Function

Two R.V.s

- Suppose we have N **samples** $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of **random variables** X, Y
- The **joint likelihood** function:
  - Case 1: X and Y are **discrete** with *pmf* $p(x,y|\theta)$
    $L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \ldots p(x^{(N)}, y^{(N)}|\theta)$
  - Case 2: X and Y are **continuous** with *pdf* $f(x,y|\theta)$
    $L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \ldots f(x^{(N)}, y^{(N)}|\theta)$
  - Case 3: Y is **discrete** with *pmf* $p(y|\beta)$ and X is **continuous** with *pdf* $f(x|y,\alpha)$
    $L(\alpha, \beta) = f(x^{(1)}|\, y^{(1)}, \alpha)\, p(y^{(1)}|\beta) \ldots f(x^{(N)}|\, y^{(N)}, \alpha)\, p(y^{(N)}|\beta)$
  - Case 4: Y is **continuous** with *pdf* $f(y|\beta)$ and X is **discrete** with *pmf* $p(x|y,\alpha)$
    $L(\alpha, \beta) = p(x^{(1)}|\, y^{(1)}, \alpha)\, f(y^{(1)}|\beta) \ldots p(x^{(N)}|\, y^{(N)}, \alpha)\, f(y^{(N)}|\beta)$

Mixed discrete/ continuous!

في حالتين كمان اللي هم Mixed discrete/continuous ... طيب افترض ان عندنا شوية داتا ..عاوزين نفكر في ال MLE شويه .. كريفيو يعني وكدا .. احنا قلنا ان ال maximizing likelihood هدفو انو بيقول انا عندي finite amount of probability mass و كل اللي هو بيحاول يعملو انو ي allocate much probability للحاجات اللي احنا لاحظناها عن الحاجات اللي احنا ملاحظنهاش ... فاحنا عندنا recipe للكلوزد فورم MLE ..

Recipe for Closed-form MLE

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

poll.mlcourse.org

MLE

Question:

Assume we have N samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ drawn from a Bernoulli(ϕ).

What is the **log-likelihood** of the data $\ell(\phi)$?

Assume $N_1 = \#$ of $(x^{(i)} = 1)$
$N_0 = \#$ of $(x^{(i)} = 0)$

Answer:

A. $l(\phi) = N_1 \log(\phi) + N_0 (1 - \log(\phi))$
B. $l(\phi) = N_1 \log(\phi) + N_0 \log(1-\phi)$
C. $l(\phi) = \log(\phi)^{N_1} + (1 - \log(\phi))^{N_0}$
D. $l(\phi) = \log(\phi)^{N_1} + \log(1-\phi)^{N_0}$
E. $l(\phi) = N_0 \log(\phi) + N_1 (1 - \log(\phi))$
F. $l(\phi) = N_0 \log(\phi) + N_1 \log(1-\phi)$
G. $l(\phi) = \log(\phi)^{N_0} + (1 - \log(\phi))^{N_1}$
H. $l(\phi) = \log(\phi)^{N_0} + \log(1-\phi)^{N_1}$
I. ~~$l(\phi)$ = the most likely answer~~

calamity.

$\boxed{\text{MLE of Bernoulli}}$

① Model $x^{(i)} \sim \text{Bernoulli}(\phi)$

② Log-likelihood: $D = \{x^{(1)}, \ldots, x^{(N)}\}$

$$\ell(\phi) = \log p(D|\phi)$$
$$= \log \prod_{i=1}^{N} p(x^{(i)}|\phi) \quad = \begin{cases} \phi & \text{if } x^{(i)} = 1 \\ 1-\phi & \text{if } x^{(i)} = 0 \end{cases} = \phi^{x^{(i)}} (1-\phi)^{(1-x^{(i)})}$$
$$= \log \prod_{i=1}^{N} \phi^{x^{(i)}} (1-\phi)^{(1-x^{(i)})}$$
$$= \log \left[\phi^{N_1} (1-\phi)^{N_0}\right]$$
$$= \underbrace{N_1 \log\phi + N_0 \log(1-\phi)}_{\log p(D|\phi)}$$

$N_1 = \#(x^{(i)} = 1)$
$N_0 = \#(x^{(i)} = 0)$

MLE

Question:

Assume we have N samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ drawn from a Bernoulli(ϕ).

What is the **derivative** of the log-likelihood $\partial \ell(\boldsymbol{\theta})/\partial\theta$?

Assume $N_1 = \#\ of\ (x^{(i)} = 1)$

$N_0 = \#\ of\ (x^{(i)} = 0)$

Answer: calamity

A. $\cancel{\partial \ell(\boldsymbol{\theta})/\partial\theta = \phi^{N_1} + (1 - \phi)^{N_0}}$

B. $\partial \ell(\boldsymbol{\theta})/\partial\theta = \phi / N_1 + (1 - \phi) / N_0$

C. $\partial \ell(\boldsymbol{\theta})/\partial\theta = N_1 / \phi\ \cancel{+}\,-\ N_0 / (1 - \phi)$

D. $\partial \ell(\boldsymbol{\theta})/\partial\theta = \log(\phi) / N_1 + \log(1 - \phi) / N_0$

E. $\partial \ell(\boldsymbol{\theta})/\partial\theta = N_1 / \log(\phi) + N_0 / \log(1 - \phi)$

② Derivative

$$\frac{\partial \ell(\phi)}{\partial \phi} = \frac{\partial}{\partial \phi}\left[N_1 \log \phi + N_0 \log(1-\phi)\right]$$

$$= \frac{N_1}{\phi} - \frac{N_0}{1-\phi}$$

③ Set to zero and solve

$$\frac{N_1}{\phi} - \frac{N_0}{1-\phi} = 0 \Rightarrow \phi^{MLE} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}$$

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

Principle of Maximum *a posteriori* (MAP) Estimation:
Choose the parameters that maximize the posterior of the parameters given the data.

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} \underbrace{\overbrace{p(\mathbf{x}^{(i)}|\boldsymbol{\theta})}^{\text{likelihood}}\overbrace{p(\boldsymbol{\theta})}^{\text{Prior}}}_{}$$

Pdf

Maximum *a posteriori* (MAP) estimate

الدكتور بيقول ان ال MLE بيحاول ي maximize p(D|phi) انما ال Max a posteriori بيحاول ي Maximize p(phi|D) ...

MAP Estimation

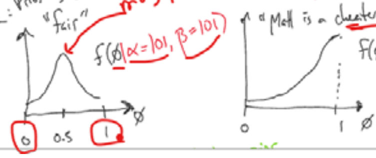
Ex: Prior Dist.

"fair" — most of the mass here

$f(\phi|\alpha=101, \beta=101)$

"Matt is a cheater"

$f(\phi|\alpha=101, \beta=1)$



source of prior is domain knowledge about the problem solving

0 0.5 1 0 1 x

General MAP: $\rightarrow \theta$

MLE: $p(D|\theta)$

MAP: $\underbrace{p(\theta|D)}_{\text{posterior}} = \underbrace{\dfrac{\overbrace{p(D|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{p(D)}}_{\text{Bayes Rule}}$ ← not a function of θ

$$\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|D) \leftarrow \text{Posterior}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \log p(\theta|D)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \log\left(\frac{p(D|\theta)p(\theta)}{\boxed{p(D)}}\right)$$

doesn't affect argmax ("remove it")

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{\log\left(p(D|\theta)p(\theta)\right)}_{\ell_{MAP}(\theta)}$$

MAP of Beta-Bernoulli

Beta Dist: $f(\phi|\alpha,\beta) = \dfrac{1}{B(\alpha,\beta)}\phi^{(\alpha-1)}(1-\phi)^{(\beta-1)}$

① Model:

$\phi \sim \text{Beta}(\alpha, \beta)$

$x^{(1)} \sim \text{Bernoulli}(\phi)$

$x^{(2)} \sim \text{Bernoulli}(\phi)$

$\vdots$

$x^{(N)} \sim \text{Bernoulli}(\phi)$

② Log-likelihood

$$\ell_{MAP}(\phi) = \log\left[p(D|\phi)\, f(\underbrace{\phi}_{\text{parameters}}|\underbrace{\alpha,\beta}_{\text{hyperparameters}})\right]$$

$$= \log\left[\left(\phi^{N_1}(1-\phi)^{N_0}\right)\left(\frac{1}{B(\alpha,\beta)}\phi^{(\alpha-1)}(1-\phi)^{(\beta-1)}\right)\right]$$

$$= \log\left[\phi^{(N_1+\alpha-1)}(1-\phi)^{(N_0+\beta-1)}\frac{1}{B(\alpha,\beta)}\right]$$

$$= (N_1+\alpha-1)\log(\phi) + (N_0+\beta-1)\log(1-\phi) - \log(B(\alpha,\beta))$$

$$= \underbrace{N_1'\log(\phi) + N_0'\log(1-\phi)}\ - \log(B(\alpha,\beta))$$

③ Derivative

$$\frac{\partial l_{MAP}(\phi)}{\partial \phi} = \frac{N_1'}{\phi} - \frac{N_0'}{1-\phi}$$

④ Set to zero and solve

$$\boxed{\phi^{MAP} = \frac{N_1'}{N_1' + N_0'} = \frac{N_1+\alpha-1}{N_1+\alpha-1+N_0+\beta-1}}$$

Real life :- عندك 10 coin flips قبل ما تلعب و بعدها تدرك لتشوف تكسب ولا لأ؟

لو انت دخلت ب prior dist. fair
بتقال 8 و 2 الي طلعو كانو noise

لو دخلت ب prior MLE
هتقول ان الدنيا مش كدة

same for (101,1)

Ex#1: Suppose $D = \{8H, 2T\}$

$$\phi^{MLE} = \frac{8}{10} = 0.8$$

pseudo counts to heads & tails

MLE

fair

cheat

Now if $\phi \sim Beta(\alpha = 101, \beta = 101)$

Prior dist.

$$\phi^{MAP_1} = \frac{8+101-1}{8+101-1+2+101-1} = \frac{108}{108+102} \approx 0.5$$

Now if $\phi \sim Beta(\alpha = 101, \beta = 1)$

$$\phi^{MAP_2} = \frac{108}{108+2} \approx 1.0$$

Ex#2: $D = \{108H, 102T\}$

$$\phi^{MLE} = \frac{108}{108+102} = \phi^{MAP_1}$$

prior Beta

اللي هو داخل ب

ودة هنا بنحطه pseudo counts

يعني كأنك عملنا قبلها فعلاً ال experiment

$\alpha = 100$ ('heads')

$\beta = 100$ ('tails')

الدكتور اتكلم علي حتة ال pseudo counts وقال ان لما انت بتلاحظ عد كبير من الداتا .. ال MLE هتبقا نفس القيمه بتاعت ال MAP رقم 1 من الصوره اللي فوق ... طيب احنا هنا كل اللي بنحاول نعملو هو function approximation و ال MLE بيدينا view مختلف شويه للتعليم .. للي بنحاول نعملو هو اننا ن Max likelihood of the data بس في ألآخر الدنيا بتوصل برضو لل function approximation ...

هنتكلم ناو علي Naïve Bayes .. هو مجرد decision function ...

Naïve Bayes

- Why are we talking about Naïve Bayes?
  - It's **just another decision function** that fits into our "big picture" recipe from last time
  - But it's our first **example of a Bayesian Network** and provides a *clearer* picture of **probabilistic learning**
  - Just like the other Bayes Nets we'll see, it **admits a closed form solution** for MLE and MAP
  - So learning is **extremely efficient** (just counting)

MLE for Naive Bayes

Data: $y \in \{0,1\}$ $\in\{H,T\}$ where real $=1$ and fake $=0$

$\vec{x} \in \{0,1\}^M$ where $M = \#$ words in vocab.

① Model:

$y \sim \text{Bernoulli}(\phi) = p(y|\phi)$

$x_1 \sim \text{Bernoulli}(\theta_{y,1}) = p(x_1|y,\theta)$

$x_2 \sim \text{Bernoulli}(\theta_{y,2}) = p(x_2|y,\theta)$

$\vdots$

$x_M \sim \text{Bernoulli}(\theta_{y,M}) = p(x_M|y,\theta)$

$\phi \in [0,1]$

$$\theta = \begin{bmatrix} \theta_{H1} & \theta_{H2} & \cdots & \theta_{HM} \\ \theta_{T1} & \theta_{T2} & \cdots & \theta_{TM} \end{bmatrix}$$

$\in [0,1]$ ← red cars ← blue cars

Tail

$$p(\underbrace{x_1, x_2, \ldots, x_M}_{\text{words}}, y|\phi,\theta) = p(y|\phi)p(x_1|y,\theta)p(x_2|y,\theta)\cdots p(x_M|y,\theta)$$

Real or fake

$$= p(y|\phi)\prod_{m=1}^{M} p(x_m|y, \theta_{H,m}, \theta_{T,m})$$

from Bernoulli

$$= \phi^y(1-\phi)^{(1-y)} \prod_{m=1}^{M} \theta_{y,m}^{x_m}(1-\theta_{y,m})^{(1-x_m)}$$

Def: two r.v.s X, Y are conditionally independent given r.v. Z written $X \perp Y | Z$ iff $p(X,Y|Z) = P(X|Z)P(Y|Z)$

Naive Bayes Assumption

$$p(\vec{x}|y) = \prod_{m=1}^{M} p(x_m|y)$$

$\Rightarrow x_q$ and x_r are cond. indep. given y