PROBABILITY, NAIVE BAYES, LOGISTIC REGRESSION

Today's lecture is advertisement for CSE 535, Probabilistic Graphical Models, and STAT 574M, Statistical Machine Learning.

THE PROBABILISTIC VIEW OF THE WORLD

If we believe data is generated at random, then if we can estimate that probability distribution, we can do "everything". Remember the Optimal Bayes Rate theorem!

This is a very useful fiction.

RULES OF PROBABILITY

Universe: $p(U) = 1 \qquad (\forall A,\ A \subset U)$

Additivity: $p(A \cup B) = p(A) + p(B)$, if $A \cap B \neq \emptyset$

Conditional Probability: $P(A \cap B) = P(B)P(A|B)$

$P(B \cap A) = P(A)P(B|A)$

That's it!

BAYES THEOREM

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

$$P(\text{SPAM} \mid \text{subject} = \text{"Get RICH"}) = \frac{P(\text{SPAM})}{P(\text{subject} = \text{"Get RICH"})} \cdot P(\text{subject} = \text{"Get RICH"} \mid \text{SPAM})$$

Lets us predict label probabilities directly!

... But we will never have enough data if we do this directly.

Instead, we make assumptions about data which lead to practical algorithms (and also lead to inductive biases)

(We often don't use Bayes's Theorem directly either)

WARMUP: COIN FLIPS

Assume coin tosses are i.i.d. $p(A \cap B) = P(B)P(A|B)$
$= P(A)P(B)$

$$P(H) = \beta$$

$$p(HHTH) = p(H)^3 \cdot p(T)$$
$$= \beta^3(1-\beta) = \beta^3 - \beta^4$$

What's the "most likely" value of β?

We often operate on logarithms instead of "naked" likelihoods — it's simply more convenient.

$$l(H, T) = \beta^H (1-\beta)^T$$

derive the maximum likelihood estimator for β.

$$ll(H,T) = H \log \beta + T \log(1-\beta)$$

$$\frac{d\,ll}{d\beta} = \frac{H}{\beta} - \frac{T}{1-\beta} = 0$$

$$H - H\beta = T\beta$$
$$H = (H+T)\beta$$
$$\beta = \frac{H}{H+T}$$

WARMUP: DIE ROLLS

Derive the MLE for K-sided die.

$$\ell(\theta) = \prod \theta_i^{c_i} \qquad \sum \theta_i = 1$$

$$\ell\ell(\theta) = \sum c_i \log \theta_i$$

$$L(\theta, \lambda) = \sum_i c_i \log \theta_i - \lambda\left(\sum \theta_i - 1\right)$$

$$\frac{\partial L}{\partial \theta_i} = \frac{c_i}{\theta_i} - \lambda = 0 \qquad \frac{c_i}{\theta_i} = \lambda \qquad \theta_i = \frac{c_i}{\lambda}$$

$$\sum \theta_i = 1 \qquad \sum \frac{c_i}{\lambda} = 1 \qquad \sum c_i = \lambda \qquad \theta_i = \frac{c_i}{\sum c_i}$$

BAYESIAN COINFLIP EXAMPLE

What's the MLE for the event "H"?

(Do you believe that?)

Bayesian Perspective: you have some prior belief about β.

If you are willing to describe that belief as a probability distribution, then the Bayes theorem gives a rule for updating your "world view", turning your prior into your "posterior".

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

PRIOR (YOU HAVE THIS)

POSTERIOR (YOU WANT THIS)

LIKELIHOOD (THIS IS USUALLY EASY TO COMPUTE)

"EVIDENCE" (THIS IS USUALLY HARD TO COMPUTE)

THINK OF THE EVIDENCE TERM AS BEING THERE TO ENSURE PROBABILITIES SUM TO 1.

$$p(\theta, \alpha, \beta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!} \quad \theta^{(\alpha-1)} (1-\theta)^{(\beta-1)}$$ (PRIOR)

(The "beta" distribution)

$$p(H \mid \theta) = \theta$$
$$p(y=1 \mid \theta) = \theta^{y}$$
$$p(y \mid \theta) = \theta^{y}(1-\theta)^{(1-y)}$$

$H \to y=1 \quad T \to y=0$

POSTERIOR:

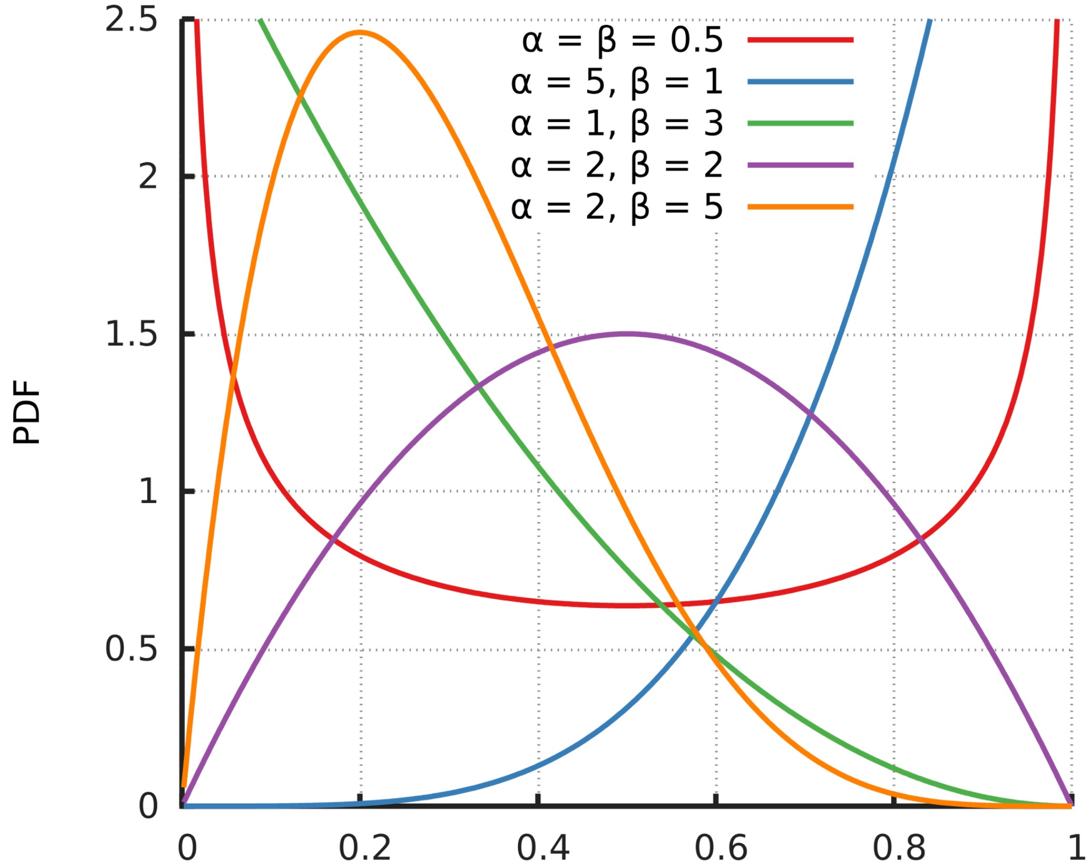
$$p(\theta, \alpha, \beta \mid y) = p(y \mid \theta, \alpha, \beta) \cdot \frac{1}{p(y)} \cdot p(\theta, \alpha, \beta)$$

$$p(\theta, \alpha, \beta \mid y) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!} \cdot \frac{1}{P(y)} \theta^{(\alpha-1)} (1-\theta)^{(\beta-1)} \cdot \theta^{y} (1-\theta)^{(1-y)}$$

$$= \frac{(\alpha+y)!}{(\alpha+y-1)!\,(\beta+1-y)!} \; \theta^{(\alpha-1+y)} \cdot (1-\theta)^{(\beta-1+(1-y))}$$

SAME SHAPE! SO WE KNOW HOW TO NORMALIZE!

α, β store <u>counts</u>!



What do we do with this?

We now have an entire distribution.

- Pick the maximum value: MAP
  ("maximum a posteriori")
  Does <u>not</u> match MLE!

MLE: $\dfrac{H}{H+T}$ MAP: $\dfrac{\alpha + H - 1}{\alpha + \beta + H + T - 2}$

- Simulate downstream using $p(\Theta)$
  - This is "the true Bayesian view"

NAIVE BAYES

Probabilities are multidimensional:

$P(A_1 \dots A_n \quad)$ depends on exponentially many relationships.

We need to assume things. If we split our PDFs in features and labels, and then assume that, conditioned on a label, features are independent, this is called the naive Bayes assumption. It's extremely naive, and extremely powerful.

$$p(x_i \mid y, x_j) = p(x_i \mid y)$$

Under Naive Bayes, we get that the general PDF for labels + features is:

$$\rightarrow \quad p((\vec{x}, y)) = p(y) \cdot \prod_i p(x_i \mid y) \quad \leftarrow$$

$$p_\theta((y, x)) = p_\theta(y) \prod_d p_\theta(x_d \mid y) \qquad \text{naive Bayes assumption} \tag{9.18}$$

$$= \left(\theta_0^{[y=+1]} (1-\theta_0)^{[y=-1]}\right) \prod_d \theta_{(y),d}^{[x_d=1]} (1-\theta_{(y),d})^{[x_d=0]} \qquad \text{model assumptions} \tag{9.19}$$

$$\hat{\theta}_0 = \frac{1}{N}\sum_n [y_n = +1]$$

$$\hat{\theta}_{(+1),d} = \frac{\sum_n [y_n = +1 \wedge x_{n,d} = 1]}{\sum_n [y_n = +1]}$$

$$\hat{\theta}_{(-1),d} = \frac{\sum_n [y_n = -1 \wedge x_{n,d} = 1]}{\sum_n [y_n = -1]}$$

PREDICTION

$$\begin{aligned} \mathrm{LLR} = \log &\left[\theta_0 \prod_d \theta_{(+1),d}^{[x_d=1]} (1-\theta_{(+1),d})^{[x_d=0]}\right] \\ &- \log\left[(1-\theta_0)\prod_d \theta_{(-1),d}^{[x_d=1]}(1-\theta_{(-1),d})^{[x_d=0]}\right] \end{aligned} \tag{9.23}$$

model assumptions

$$\begin{aligned} &= \log\theta_0 - \log(1-\theta_0) + \sum_d [x_d = 1]\left(\log\theta_{(+1),d} - \log\theta_{(-1),d}\right) \\ &\quad + \sum_d [x_d = 0]\left(\log(1-\theta_{(+1),d}) - \log(1-\theta_{(-1),d})\right) \end{aligned} \tag{9.24}$$

take logs and rearrange

$$= \sum_d x_d \log\frac{\theta_{(+1),d}}{\theta_{(-1),d}} + \sum_d (1-x_d)\log\frac{1-\theta_{(+1),d}}{1-\theta_{(-1),d}} + \log\frac{\theta_0}{1-\theta_0} \tag{9.25}$$

simplify log terms

$$= \sum_d x_d\left[\log\frac{\theta_{(+1),d}}{\theta_{(-1),d}} - \log\frac{1-\theta_{(+1),d}}{1-\theta_{(-1),d}}\right] + \sum_d \log\frac{1-\theta_{(+1),d}}{1-\theta_{(-1),d}} + \log\frac{\theta_0}{1-\theta_0} \tag{9.26}$$

group x-terms

$$= \boldsymbol{x}\cdot\boldsymbol{w} + b \tag{9.27}$$

$$w_d = \log\frac{\theta_{(+1),d}(1-\theta_{(-1),d})}{\theta_{(-1),d}(1-\theta_{(+1),d})}\ , \quad b = \sum_d \log\frac{1-\theta_{(+1),d}}{1-\theta_{(-1),d}} + \log\frac{\theta_0}{1-\theta_0} \tag{9.28}$$

LINEAR MODEL!