Binary Density Estimation

CPSC 440/550: Advanced Machine Learning

`cs.ubc.ca/~dsuth/440/24w2`

University of British Columbia, on unceded Musqueam land

2024-25 Winter Term 2 (Jan–Apr 2025)

Admin

admin

- Sign up for Piazza from the link on `cs.ubc.ca/~dsuth/440`
- Lecture recordings are linked from Piazza

- CBTF quiz booking should be available by the end of this week
- Will post instructions on Piazza once it's available

- Again, I expect everyone to get in off the waitlist
  - But it'll take a bit to confirm and sort through everything

- Assignment 1 will be out tonight
- If you're on the waitlist (and want to join the class), **do the assignment**

- Office hours starting next week – will link calendar from Piazza

Last time: binary density estimation

- Density estimation: going from data → probability model
- Inference: "doing things" with a probability model
  - Computing probabilities of "derived events"
  - Computing likelihoods
  - Finding the mode
  - Sampling

- Bernoulli distribution: simple parameterized probability model for binary data
- If $X \sim \mathrm{Bern}(\theta)$, then for $x \in \{0, 1\}$ we have

$$\Pr(X = x \mid \theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases} = \theta^{\mathbb{1}(x=1)}(1-\theta)^{\mathbb{1}(x=0)} = \theta^x(1-\theta)^{1-x}$$

- Also write this as $p(x \mid \theta)$ or even $p(x)$, if context is clear

Outline

1. Maximum likelihood estimation (MLE)

2. MAP estimation

MLE: binary density estimation

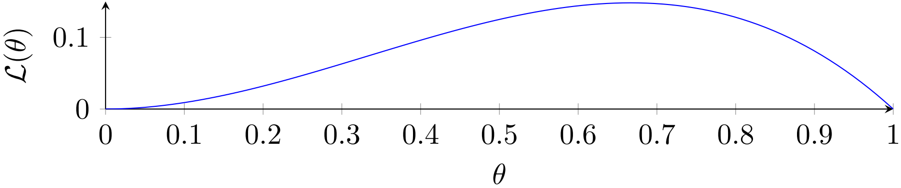
- We know how to use a Bernoulli model (inference) for a bunch of tasks
- How can we train a Bernoulli model (learning) from data?

$$\mathbf{X} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \xrightarrow{\text{MLE}} \quad \theta = 0.4$$

  - Recall $\mathbf{X}$ collects the data points $x^{(1)}, \ldots, x^{(n)}$
  - We assume these are iid samples from a random variable X
- Classic way: maximum likelihood estimation (MLE)

The likelihood function

- The likelihood function is a function from parameters θ to the probability (density) of the data under those parameters
  - $\mathcal{L}(\theta) = p(\mathbf{X} \mid \theta)$, which for Bernoullis we saw is $\theta^{n_1}(1-\theta)^{n_0}$
- Here's the likelihood for $\mathbf{X} = (1, 0, 1)$, i.e. $\theta^2(1-\theta)$:



  - $\mathcal{L}(0.5) = p(1, 0, 1 \mid \theta = 0.5) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.125$
  - $\mathcal{L}(0.75) = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \approx 0.14$: $\mathbf{X}$ is more likely for $\theta = 0.75$ than $\theta = 0.5$
  - $\mathcal{L}(0) = 0 = \mathcal{L}(1)$: $\mathbf{X}$ is impossible for $\theta = 0$ or 1, since we have some 1s and some 0s
  - Maximum is at $\theta = 2/3$ – back to this in a second
- Likelihood is not a distribution over θ, i.e. $\int \mathcal{L}(\theta) \, \mathrm{d}\theta \neq 1$
  - We do have $\int p(\mathbf{X} \mid \theta) \, \mathrm{d}\mathbf{X} = 1$, but that's not really relevant if we only have one $\mathbf{X}$

Maximizing the likelihood

- Maximum likelihood estimation (MLE): pick the θ with the highest likelihood
  - "Find the parameters θ where the data $\mathbf{X}$ would have been most likely to be seen"
- For Bernoullis, the MLE is $\hat{\theta} = \dfrac{n_1}{n} = \dfrac{n_1}{n_1 + n_0}$
  - "If you flip a coin 50 times and get 23 heads, guess that $\Pr(\texttt{heads}) = \frac{23}{50}$"
  - Code: `theta = np.mean(X)` takes $\mathcal{O}(n)$ time
- Let's derive this result
  - It's going to seem overly complicated for this really simple result
  - But the steps we use will be applicable to much harder situations

MLE for Bernoullis

- Notationally, we can write maximizing the likelihood as

$$\hat{\theta} \in \underset{\theta}{\arg\max}\, \mathcal{L}(\theta) = \underset{\theta}{\arg\max}\ \theta^{n_1}(1-\theta)^{n_0}$$

- $\arg\max_x f(x)$ means "the set of x that maximize f": might be more than one!
- Usually, instead of maximizing the likelihood we maximize the log-likelihood
  - Same solution set, since if $\alpha > \beta$ then $\log \alpha > \log \beta$ ($\log$ is strictly monotonic)
    - See "Max and Argmax" notes from the course site
  - Usually easier mathematically (also numerically much more stable)

$$\hat{\theta} \in \underset{\theta}{\arg\max}\ n_1 \log(\theta) + n_0 \log(1-\theta)$$

- The maximum will have a zero derivative:

$$0 = \frac{n_1}{\theta} - \frac{n_0}{1-\theta}$$

- and so $n_1(1-\theta) = n_0\theta$ or $n_1 = \underbrace{(n_0+n_1)}_{n}\theta$ or $\theta = \dfrac{n_1}{n}$

MLE for Bernoullis

- We're looking for

$$\hat{\theta} \in \arg\max_{\theta} \log \mathcal{L}(\theta) = \arg\max_{\theta}\ n_1 \log(\theta) + n_0 \log(1-\theta)$$

- Derivative of $n_1 \log(\theta) + n_0 \log(1-\theta)$ is zero only if $\theta = \frac{n_1}{n_0+n_1} = \frac{n_1}{n}$
- But is this actually a maximum?
- Yes: it's a concave function (second derivative is negative): $-\frac{n_1}{\theta^2} - \frac{n_0}{(1-\theta)^2} \leq 0$

- What if $n_1 = 0$ or $n_0 = 0$? Then we just divided by zero!
- $\log(0) = -\infty$ makes things complicated; go back to plain likelihood $\theta^{n_1}(1-\theta)^{n_0}$
- If $(n_1 = 0,\ n_0 > 0)$, find $\theta = 0$; if $(n_1 > 0,\ n_0 = 0)$, get $\theta = 1$
  - So same n_1/n formula still works

MLE for binary data estimation

- Given iid binary data $\mathbf{X}$, we can train/learn a probability model with MLE:

$$\mathbf{X} \quad \xrightarrow{\text{MLE}} \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

- Given this $\mathrm{Bern}(\hat{\theta})$ model, can then ask inference questions
    - "If I eat lunch with three randomly selected UBC students, what's the probability any of them are COVID-positive?"
        - One minus the probability none of them are: $1 - (1-\theta)^3 \approx (1 - (1-\hat{\theta})^3)$

Outline

1. Maximum likelihood estimation (MLE)
2. MAP estimation

Problems with MLE

- Often (including here), the MLE is asymptotically optimal as $n \to \infty$
  - In particular, if we see $X \sim \mathrm{Bern}(\theta^*)$, then $\hat{\theta}$ converges to the true θ^* as $n \to \infty$
  - These kinds of properties are covered in honours/grad stat classes

- But for small n, it can do really bad things
  - Before we considered $x^{(1)} = 1, x^{(2)} = 0, x^{(3)} = 1$, with $\hat{\theta}_{\mathrm{MLE}} \approx 0.67$
  - If we see an $x^{(4)} = 1$, we get an MLE of 0.75
  - If we see an $x^{(4)} = 0$, get an MLE of 0.5
  - If you get an "unlucky" $\mathbf{X}$, the MLE might be really bad

- For Bernoullis, this sensitivity decreases quickly with n
- But for more complex models, the MLE can tend to overfit

Problems with MLE

- Imagine instead we'd seen a (barely-different) dataset, $x^{(1)} = 1$, $x^{(2)} = 1$, $x^{(3)} = 1$
- Then the MLE is $\hat{\theta} = 1$

- Now imagine we see a test dataset with a 0 in it
- Our likelihood of that test dataset is zero, because $1 - \hat{\theta} = 0$
  - Serious overfitting to this small dataset
  - If your drug works for everyone in a trial of three people, does it *always* work?

- Common solution (340 does this for Naive Bayes): Laplace smoothing

$$\hat{\theta}_{\text{Lap}} = \frac{n_1 + 1}{(n_1 + 1) + (n_0 + 1)} = \frac{n_1 + 1}{n + 2}$$

- MLE for a dataset with an extra "imaginary" 0 and 1 in it; avoids zero counts
- This is a special case of MAP estimation

Following a MAP

- In MLE we maximize the probability of the data given the parameters:

$$\hat{\theta} \in \arg\max_{\theta} p(\mathbf{X} \mid \theta)$$

- "Find the θ that makes $\mathbf{X}$ have the highest probability given θ"
- But. . . this is kind of weird
- Data could be most likely for a really weird θ: get overfitting
  - If θ allows highly-complex models, could be one that just memorizes the data exactly

- What we really want is the "best" θ
- "After seeing the data $\mathbf{X}$, which θ is most likely?"

$$\hat{\theta} \in \arg\max_{\theta} p(\theta \mid \mathbf{X})$$

- This is called maximum a posteriori (MAP) estimation

Probability review (MAKE SURE YOU KNOW ALL OF THIS)

review

- Product rule: $\Pr(A \cap B) = \Pr(A \mid B)\Pr(B)$
    - Rearrange into conditional probability formula: $\Pr(A \mid B) = \Pr(A \cap B)/\Pr(B)$
    - Order doesn't matter for joints: $\Pr(A \cap B) = \Pr(B \cap)$
    - Using twice, get Bayes rule: $\Pr(A \mid B) = \Pr(B \mid A)\Pr(A)/\Pr(B)$
        - Flips order of conditionals, depending on the marginals $\Pr(A)$ and $\Pr(B)$
- Marginalization rule:
    - If X is discrete: $\Pr(A) = \sum_x \Pr\left(A \cap (X = x)\right)$
    - If X is continuous: $\Pr(A) = \int p\left(A \cap (X = x)\right) \mathrm{d}x$
- These two rules are close friends:

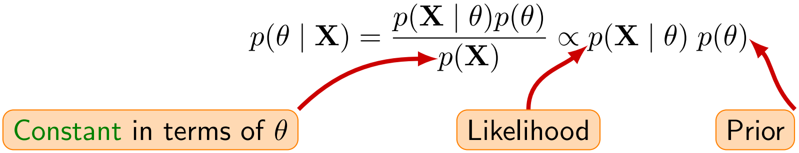
$$p(a) = \sum_b p(a, b) = \sum_b p(a \mid b) p(b); \quad p(a \mid b) = \frac{p(b \mid a) p(a)}{p(b)} = \frac{p(b \mid a) p(a)}{\sum_{a'} p(b \mid a') p(a')}$$

- Still work if you condition everything:
    - $p(a, b \mid c) = p(a \mid b, c) p(b \mid c)$ and $p(a \mid c) = \sum_b p(a, b| c)$

- See probability notes on the course site if you need them (catch up quick!)

Maximum a Posteriori (MAP) estimation

- Posterior probability is "what we believe *after* seeing the data": $p(\theta \mid \mathbf{X})$
- Using Bayes rule,



$$p(\theta \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \theta) p(\theta)}{p(\mathbf{X})} \propto p(\mathbf{X} \mid \theta)\, p(\theta)$$

Constant in terms of θ

Likelihood

Prior

- To use this, we need a prior distribution for θ
  - What we believe about θ *before* seeing the data
  - If we're flipping coins: might want $p(\theta)$ higher for values close to/exactly equal to $\frac{1}{2}$
  - For COVID, maybe a separate study estimated Lower Mainland rate at 0.04
    - Then could use a prior that prefers θ not too different from that number
  - In CPSC 340, priors on linear models' weights correspond to regularizers
    - Choose smaller $p(\theta)$ for models more likely to overfit

MAP for Bernoulli with a discrete prior

- Consider $x^{(1)} = 1$, $x^{(2)} = 1$, $x^{(3)} = 0$, where MLE is $\frac{2}{3}$

Using a prior that looks like	Gives posterior proportional to
$\Pr(\theta = 0) = 0.05$	$\Pr(\theta = 0 \mid \mathbf{X}) \propto (0 \cdot 0 \cdot 1) \cdot 0.05 = 0$
$\Pr(\theta = 0.25) = 0.2$	$\Pr(\theta = 0.25 \mid \mathbf{X}) \propto (0.25 \cdot 0.25 \cdot 0.75) \cdot 0.2 \approx 0.01$
$\Pr(\theta = 0.5) = 0.5$	$\Pr(\theta = 0.5 \mid \mathbf{X}) \propto (0.5 \cdot 0.5 \cdot 0.5) \cdot 0.5 \approx 0.06$
$\Pr(\theta = 0.75) = 0.2$	$\Pr(\theta = 0.75 \mid \mathbf{X}) \propto (0.75 \cdot 0.75 \cdot 0.25) \cdot 0.2 \approx 0.03$
$\Pr(\theta = 1) = 0.05$	$\Pr(\theta = 1 \mid \mathbf{X}) \propto (1 \cdot 1 \cdot 0) \cdot 0.05 = 0$

- So our MAP estimate is $\hat{\theta} = 0.5$
  - . . . using this choice of prior, which favours a fair coin
- Notice that $p(\mathbf{X})$ didn't matter: it's the same for all θ

Digression: proportional-to ($\propto$) notation

- In math, the notation $f(\theta) \propto g(\theta)$ means "there is some $\kappa > 0$ such that $f(\theta) = \kappa g(\theta)$ for all θ"
- There are many possible κ: we have both $10\theta^2 \propto \theta^2$ and $\sqrt{\pi}\theta^2 \propto \theta^2$

- For probability distributions, if $p \propto g$, the constant κ is unique

- This is because we know that probability distributions sum/integrate to 1:
- Say θ is discrete, and $p(\theta) = \kappa g(\theta) \propto g(\theta)$
  - We know that $\sum_\theta p(\theta) = 1$, so $\sum_\theta \kappa g(\theta) = 1$: thus $\kappa = 1/\left(\sum_\theta g(\theta)\right)$
  - Plugging back in, this means $p(\theta) = \frac{g(\theta)}{\sum_{\theta'} g(\theta')}$

- Plugging in on the previous slide, we could find that e.g.

$$\Pr(\theta = 0.5 \mid \mathbf{X}) \approx \frac{0.06}{0 + 0.01 + 0.06 + 0.03 + 0} \approx 60\%$$

- Using $\propto$ can make our life a lot easier!

Continuous distributions

review

- Recall that θ could be any number between 0 and 1
- But our previous prior only allowed $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$
- Instead, it'd be nicer to allow any value of θ from $[0, 1]$

- Usually want a continuous distribution
- Convenient to work with their probability density function (pdf)
  - A function $p(\theta)$ with $p(\theta) \geq 0$ and $\int_{-\infty}^{\infty} p(\theta) \mathrm{d}\theta = 1$
    - Note: can have $p(\theta) > 1$ for some θ!
  - Get probabilities by integrating over a range: $\Pr(0.45 \leq \theta \leq 0.55) = \int_{0.45}^{0.55} p(\theta) \, \mathrm{d}\theta$
  - Probability of any individual θ is 0: $\Pr(\theta = 0.5) = \int_{0.5}^{0.5} p(\theta) \, \mathrm{d}\theta = 0$

- Note that if $p \propto g$, $1 = \int p(\theta) \mathrm{d}\theta = \kappa \int g(\theta) \mathrm{d}\theta$
  - Proportionality constant is still unique, $p(\theta) = g(\theta) / \int g(\theta') \mathrm{d}\theta'$

Continuous posteriors

- Recall the posterior, likelihood, prior are related as

$$p(\theta \mid \mathbf{X}) \propto p(\mathbf{X} \mid \theta)\, p(\theta)$$

- If we have a continuous prior on θ, $p(\theta)$ is a probability *density*
- But even so, for binary $\mathbf{X}$, likelihood $p(\mathbf{X} \mid \theta)$ is a probability:

$$p(\mathbf{X} \mid \theta) = \Pr(X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)} \mid \theta)$$

  - Later, for continuous X, likelihood will also be a density function
- $p(\theta \mid \mathbf{X})$ is also a posterior density

What prior to use for Bernoulli?

- Want a continuous distribution on $[0,1]$ that works well with a Bernoulli likelihood
- Most common choice is the beta distribution:

$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0 \leq \theta \leq 1, \alpha > 0, \beta > 0$$

  - Density is 0 if $\theta \notin [0,1]$
  - Looks like a Bernoulli likelihood, with $(\alpha-1)$ ones and $(\beta-1)$ zeroes
  - But a key difference: the argument is θ, not α or β
  - Probability distribution over $\theta \in [0,1]$ – “probability over probabilities”

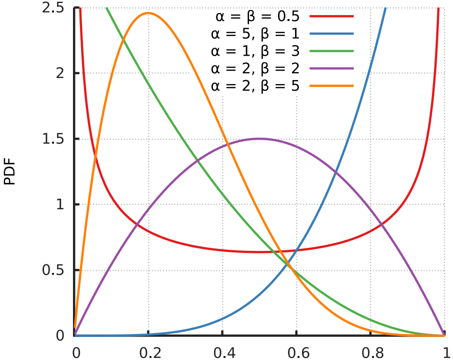
- We know what's hidden in the $\propto$ sign:

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathrm{d}\theta}$$

Beta function $B(\alpha, \beta)$

Beta distribution

- Beta distribution can take many shapes for different α and β: animation



https://en.wikipedia.org/wiki/File:Beta_distribution_pdf.svg

- Why such a popular choice? Partial reason: it's pretty flexible
  - Can prefer 0.5, 0, 0.23561, towards "0 or 1", can be uniform ($\alpha = \beta = 1$), ...
  - Can't bias towards "0.25 or 0.75", can't say "half the time it'll be *exactly* 0.5", ...

Beta-Bernoulli model

- Beta is "flexible enough," but mostly posterior and MAP have really simple forms
- Posterior when $\theta \sim \mathrm{Beta}(\alpha, \beta)$, $X \sim \mathrm{Bern}(\theta)$:

$$\begin{aligned} p(\theta \mid \mathbf{X}, \alpha, \beta) &\propto p(\mathbf{X} \mid \theta, \alpha, \beta)\; p(\theta \mid \alpha, \beta) = p(\mathbf{X} \mid \theta) p(\theta \mid \alpha, \beta) \\ &\propto \theta^{n_1} (1-\theta)^{n_0}\; \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= \theta^{(n_1+\alpha)-1} (1-\theta)^{(n_0+\beta)-1} \end{aligned}$$

which is another beta distribution! $(\theta \mid \mathbf{X}, \alpha, \beta) \sim \mathrm{Beta}(\alpha + n_1, \beta + n_0)$

- Why does it have to be a beta? Because $\propto$ is unique
  - If $p(t) \propto t^{\tilde{\alpha}-1}(1-t)^{\tilde{\beta}-1}$, we necessarily have $t \sim \mathrm{Beta}(\tilde{\alpha}, \tilde{\beta})$
  - Make sure this makes sense to you!

MAP in the Beta-Bernoulli model

- The posterior with a Bernoulli likelihood and beta prior is beta
- That is, with $\tilde{\alpha} = n_1 + \alpha$, $\tilde{\beta} = n_0 + \beta$,

$$p(\theta \mid \mathbf{X}, \alpha, \beta) = \frac{\theta^{\tilde{\alpha}-1}(1-\theta)^{\tilde{\beta}-1}}{B(\tilde{\alpha}, \tilde{\beta})}$$

- Taking the log and setting the derivative to zero gives

$$\theta = \frac{\tilde{\alpha}-1}{\tilde{\alpha}+\tilde{\beta}-2} = \frac{n_1+\alpha-1}{n+\alpha+\beta-2} \quad \text{or} \quad \theta \in \{0, 1\}$$

- If $\tilde{\alpha} > 1$, $\tilde{\beta} > 1$ (always true if $n_0, n_1 \geq 1$), then MAP is first expression above
  - If $\alpha = 1$, $\beta = 1$ (a uniform prior), we get the MLE
  - If $\alpha = \beta = 2$ (mild preference towards $1/2$), we get Laplace smoothing
  - If $\alpha = \beta > 2$, we bias more strongly towards $\hat{\theta} = 0.5$ than Laplace smoothing
  - If $\alpha = \beta < 1$, we bias away from 1/2 (towards either 0 or 1)
  - If $\alpha > \beta$, we bias towards 1
  - As $n \to \infty$, the prior stops mattering and MAP → MLE
    - But using a prior means we behave better when we have relatively small n

Existence of MAP estimate under beta prior

- Our MAP estimate for $\mathrm{Beta}(\alpha, \beta)$ prior and Bernoulli likelihood was

$$\hat{\theta} = \frac{n_1 + \alpha - 1}{(n_1 + \alpha - 1) + (n_0 + \beta - 1)}$$

  - We assumed that $n_1 + \alpha > 1$, $n_0 + \beta > 1$

- But what what if we don't have these?
- By checking likelihood, get pretty quickly that:
  - If $n_1 + \alpha > 1$ and $n_0 + \beta \leq 1$, $\hat{\theta} = 1$
  - If $n_1 + \alpha \leq 1$ and $n_0 + \beta > 1$, $\hat{\theta} = 0$
  - If $n_1 + \alpha < 1$ and $n_0 + \beta < 1$, density is infinite at both $\hat{\theta} = 0$ and $\hat{\theta} = 1$
  - If $n_1 + \alpha = 1$ and $n_0 + \beta = 1$, anything in $[0, 1]$ works

Hyper-parameters and (cross)-validation

review

- We call the parameters of the prior, α and β, the hyper-parameters
  - Parameters that "affect the complexity of the model"
    - 340 examples: degree of a polynomial, depth of a decision tree, neural network architecture, regularization weight, number of rounds of gradient boosting
  - Also anything hard to fit with your learning algorithm, e.g. gradient descent step size
- Trying to fit α and β based on training likelihood doesn't work: would just become MLE by making $\alpha, \beta \to 1$
- Default 340-type approach: use a validation set (or cross-validation)
  - Split $\mathbf{X}$ into "training" and "validation" sets
  - For different values of α and β:
    - Find the MAP on the training set, evaluate its validation likelihood
  - Pick the hyper-parameters with highest validation likelihood
    - Approximates maximizing the held-out generalization error on totally-new data
- 340 covers many things that can go wrong, like overfitting to the validation set
  - Happens all the time, including in UBC PhD theses and in top conferences!
- CPSC 532D covers this more mathematically :)

Summary

- Maximum likelihood estimation (MLE):
  - Estimates θ by finding the setting that maximizes the data likelihood, $p(\mathbf{X} \mid \theta)$
  - For Bernoulli, just $\hat{\theta} = (\text{number of } 1\text{s})/(\text{number of examples})$
- Maximum a posteriori (MAP) estimation:
  - Maximizes posterior probability of parameters given data
  - Can avoid bad behaviour of MLE, but requires choosing a prior
- Probability review: product rule, marginalization, Bayes rule, α for probabilities
- Beta distribution: "cooperates well" with Bernoulli likelihood

- Next time: everything(ish) from 340 but with probabilities