CPSC 440: Machine Learning

MAP Estimation

Winter 2022

Last Time: Bernoulli Distribution MLE

Wait list: Monday
Auditors: Wednesday
Assignment 1: Friday

- The Bernoulli distribution for binary variables: $p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$
- We talked about difference inference tasks in Bernoulli models:
  – Compute likelihood of data, $p(x^1, x^2,\dots,x^n \mid \theta)$.
  – Compute decoding, $\text{argmax}_x \{p(x \mid \theta\})$.
  – Generate samples $\tilde{x}$ from $p(x \mid \theta)$.
- We discussed learning with maximum likelihood estimation (MLE).
  – Find a $\hat{\theta}$ in $\text{argmax}_\theta\{p(x^1, x^2,\dots,x^n \mid \theta)\}$.
  – Equivalent to finding $\hat{\theta}$ in $\text{argmax}_\theta\{\log(p(x^1, x^2,\dots,x^n \mid \theta))\}$ ("log-likelihood").
- For Bernoulli, equating derivative with respect to θ to 0 gives:
  – $\hat{\theta} = n_1/n$ (proportion of examples that are "1").

Derivation MLE for Bernoulli

- We showed log-likelihood derivative is zero for θ = $n_1/(n_1+n_0)$.
  - Or $\theta = n_1/n$, since $n_1+n_0=n$.
- We still need to convince ourselves this is a maximum:
  - You can verify that the second derivative of log-likelihood is negative.
    - So the function is "curved downwards" and this is a maximum.
- What about if $n_1=0$ or $n_0=0$?
  - In these cases we would get a "divide by zero" in our derivation.
  - If $n_1=0$ then MLE is $\theta = 0$ and if $n_0=0$ then MLE is $\theta = 1$.
    - Can show that likelihood is increasing as it approaches 0/1 in these cases.
    - So the formula $\theta = n_1/n$ still works.

Learning Task: Computing MLE

- Computing MLE for Bernoulli in code given data ‘X’:

Version 1:

```
n1 = sum(X)
n0 = n - n0
θ = n1/(n1 + n0)
```

Version 2:

```
θ = sum(X)/n
```

- Cost: O(n).
  - You need to sum up the ‘n’ values (there is a “for” loop hidden inside “sum(X)”).

- You can then use this θ value for inference:
  - Compute likelihood of test data.
  - Compute expected number of samples before first 1.
  - Compute probability of seeing at least three 1 values in 10 samples.

Next Topic: MAP Estimation

Problems with MLE

- In most settings, MLE is optimal as ‘n’ goes to ∞.
  - It converges to the true parameter(s).
    - This is called “asymptotic consistency” (covered in honours/grad stats classes).

- However, it can be very sensitive for small ‘n’:
  - Consider our example where $x^1=1$, $x^2=1$, $x^3=0$, and MLE was 0.67.
  - If $x^4=1$, then MLE goes up to 0.75.
  - If $x^4=0$, then MLE goes down to 0.5.
    - If you get “unlucky” with your samples, the MLE might be really bad.

- For Bernoullis, this sensitivity goes away quickly as we increase ‘n’.
  - But for more complicated models, MLE tends to lead to overfitting.

Problems with MLE

- Consider a different dataset consisting of $x^1=0$, $x^2=0$, $x^3=0$.
  - In this case the MLE is $\theta = 0$.
    - It assigns zero probability to events that do not occur in training data.

- Causes problems if we have a '1' in test data:
  - Then likelihood of entire test set is 0, since: $p(\hat{X} \mid \theta) = \theta^{n_1}(1-\theta)^{n_0} = 0^{n_1} 1^{n_0} = 0$ $\quad n_1 > 0$
    - A case of overfitting to the training data.
    - If you have no COVD-19 cases in your sample, does that mean there are none in population?

- It is common to add Laplace smoothing to the estimator:

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

  - MLE for a dataset with an extra "imaginary" '1' and '0' in data.
    - This is a special case of "MAP estimation".

MLE and MAP Estimation

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

$$\hat{\theta} \in \underset{\theta}{\mathrm{argmax}} \{ p(X \mid \theta) \}$$

- But I find this weird:
  - "Find the θ that makes 'X' have the highest probability given θ."
  - Get overfitting because data could be likely for an unlikely θ.
    - For example, a complex model that overfits by memorizing the data.

- What we really want if we are trying to find the "best" θ:
  - "Find the θ that has the highest probability given the data 'X'."

$$\hat{\theta} \in \underset{\theta}{\mathrm{argmax}} \{ p(\theta \mid X) \}$$

reversed

  - This is called MAP estimation ("maximum a posteriori").

Digression: Super-Quick "Probability Rule" Review

- Product rule: p(a,b) = p(a | b)p(b).
  - Re-arrange to get conditional probability formula: p(a | b) = p(a,b)/p(b).
  - Order dot not matter in joint probabilities: p(a,b) = p(b, a).
  - Use product rule twice to get Bayes rule: p(a | b) = p(b | a)p(a)/p(b).
    - Conditional in terms of "reverse" conditional, and the "marginals" p(b) and p(a).

- Marginalization rule ("summing or integrating over a variable"):
  - Variable 'b' with discrete domain: p(a) = $\sum_b p(a,b)$.
  - Variable 'b' with continuous domain 'b': p(a) = $\int p(a,b)db$.

- These two rules are good friends and usually appear together:
  - p(a) = $\sum_b p(a,b) = \sum_b p(a|b)p(b)$.
  - p(a | b) = $\frac{p(b\,|a)p(b)}{p(a)} = \frac{p(b\,|a)p(b)}{\sum_b p(b\,|a)p(b)}$ (some people call this "Bayes rule").

- Rules still work if you add extra "conditioning" on the right:
  - p(a,b | c) = p(a | b, c)p(b | c).
  - p(a | c) = $\sum_b p(a,b\,|c)$.

MEMORIZE EVERYTHING ON THIS SLIDE

Maximum a Posteriori (MAP) Estimation

- Maximum a posteriori (MAP) estimate maximizes posterior probability:

$$\hat{\theta} \in \underset{\theta}{\operatorname{argmax}} \{ \underbrace{p(\theta \mid X)}_{\text{"posterior"}} \}$$

  – I would argue that this is what we want: the probability of θ given our data.

- MLE and MAP are connected by Bayes rule:

$$\underbrace{p(\theta \mid X)}_{\text{(posterior)}} = \frac{p(X, \theta)}{p(X)} = \frac{p(X \mid \theta)p(\theta)}{p(X)} \propto \underbrace{p(X \mid \theta)}_{\text{(likelihood)}} \underbrace{p(\theta)}_{\text{(prior)}}$$

"proportional to"

definition of conditional probability

"product rule" $p(a,b) = p(a|b)p(b)$

constant that does not depend on θ

  – So posterior is proportional the likelihood $p(X|\theta)$ times the prior $p(\theta)$.
    - See "probability" notes on course webpage if equalities above aren't obvious (you need catch up fast).

The prior

- The prior $p(\theta)$ can encode our preference for different parameters.
  - If we are flipping coins, we might think $p(\theta)$ is higher for values close to ½.
    - We could make it really high for the exact value ½.
  - In COVID-19 example, we might make $p(\theta)$ higher for values close to 0.05.
    - Because, for example, we estimated a value of 0.05 from a similar population.
  - In CPSC 340, you learned that priors correspond to regularizers.
    - You often choose $p(\theta)$ to be lower for values that are likely to overfit.

- Laplace smoothing corresponds to a particular $p(\theta)$.
  - We will show this shortly.

MAP Estimation for Bernoulli with Discrete Prior

- Consider our example where $x^1=1$, $x^2=1$, $x^3=0$ (and MLE was 0.67).
- Consider using a prior of:
  - $p(\theta = 0.00) = 0.05$
  - $p(\theta = 0.25) = 0.2$
  - $p(\theta = 0.50) = 0.5$
  - $p(\theta = 0.75) = 0.2$
  - $p(\theta = 1.00) = 0.05$

  Posterior values are proportional to:
  - $p(\theta = 0.00 \mid X) \propto (0*0*1)*.05 = 0$
  - $p(\theta = 0.25 \mid X) \propto (.25*.25*.75)*.2 \approx 0.01$
  - $p(\theta = 0.50 \mid X) \propto (.5*.5*.5)*.5 \approx 0.06$
  - $p(\theta = 0.75 \mid X) \propto (.75*.75*.25)*.2 \approx 0.03$
  - $p(\theta = 1.00 \mid X) \propto (1*1*0)*.05 = 0$
- So our MAP estimate is $\theta = 0.5$.
  - Based on our prior "guesses for θ", we think this is a fair coin.
    - Notice that we don't need p(X) in our calculations (since it's the same for all θ).

Digression: "Proportional to" ($\propto$) Notation

- In math, the notation $f(\theta) \propto g(\theta)$
  means that $f(\theta) = \kappa g(\theta)$ for some number κ (for all θ).
  - But κ may not be known and/or may not be unique.
    - For example, $f(\theta) \propto \theta^2$ for both $f(\theta)= 10\theta^2$ and $f(\theta) = -50\theta^2$.

- For discrete probabilities, the constant κ is positive and unique.
  - This is because probabilities are non-negative and sum to 1.

- Consider a discrete variable 'θ' with $p(\theta)= \kappa g(\theta) \propto g(\theta)$:
  - Since $\sum_{\theta'} p(\theta') = 1$, we have $\sum_{\theta'} \kappa g(\theta') = 1$.
    - Solving for κ gives: $\kappa = \frac{1}{\sum_{\theta'} g(\theta')}$.
  - Using this value for κ we have $p(\theta) = \kappa g(\theta) = \frac{g(\theta)}{\sum_{\theta'} g(\theta')}$.
  - You can use this trick to get posterior probabilities on last slide: $p(\theta=0.5 \mid X) = \frac{0.06}{0+0.01+0.06+0.03+0}$

Values the posterior was proportional to.

Digression²: "Probability" vs. "Probability Density"

- Recall that the value θ can be any number between 0 and 1.
  - Instead of putting non-zero probability on a finite number of possible θ values, we could treat θ as a continuous random variable (to allow $\theta = 0.3452$).

- For continuous variables, we use a probability density function (PDF):
  - Function 'p' that is non-negative and integrates to 1 over domain:

$$p(\theta) \geq 0 \quad \text{for all } \theta, \quad \text{and} \quad \int_{-\infty}^{\infty} p(\theta)\,d\theta = 1$$

- We get probabilities from the PDF by integrating over ranges:

$$prob(0.45 \leq \theta \leq 0.55) = \int_{0.45}^{0.55} p(\theta)\,d\theta$$

  - If the PDF is continuous, probability of an individual θ is 0: $prob(\theta = 0.5) = \int_{0.5}^{0.5} p(\theta)\,d\theta = 0$

Digression^2: "Probability" vs. "Probability Density"

- Recall the relationship between posterior, likelihood, and prior:

$$\underbrace{p(\theta \mid X)}_{\text{(posterior)}} \propto \underbrace{p(X \mid \theta)}_{\text{(likelihood)}} \underbrace{p(\theta)}_{\text{(prior)}}$$

- What are these 'p' functions in discrete and continuous case?
  - If θ is discrete: prior and posterior 'p' functions are probabilities.
  - If θ is continuous: prior and posterior 'p' functions are PDFs.
    - So p(θ) is not the "probability of θ", but the "probability density of θ".

- With our binary 'X' values, likelihood p(X | θ) is a probability.
  - But when we later talk about continuous 'X', likelihood will be a PDF.
- Important: I'm really sloppy about this! (Most ML people are!)
  - I will usually say "probability of θ" for p(θ), even for continuous θ.

Digression: "Proportional to" (∝) Notation

- Consider a continuous variable θ with PDF $p(\theta) = \kappa g(\theta) \propto g(\theta)$:
  - Since $\int_{\theta'} p(\theta')d\theta' = 1$, we have $\int_{\theta'} \kappa g(\theta')d\theta' = 1$.
    - Solving for κ gives: $\kappa = \frac{1}{\int_{\theta'} g(\theta')d\theta'}$.
  - So we have $p(\theta) = \frac{g(\theta)}{\int_{\theta'} g(\theta')d\theta'}$.

- For continuous θ in MAP estimation, we have $p(\theta \mid X) \propto p(X \mid \theta)p(\theta)$,
  - So we have $p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\underbrace{\int_{\theta'} p(X \mid \theta')p(\theta')d\theta'}}$ $\to = p(X)$ by "marginalization rule": $p(a) = \sum_b p(a,b)$ (discrete) or $p(a) = \int_b p(a,b)db$ (continuous)

- You should memorize these "digression" slides.
  - Knowing how to use "∝" simplifies a lot of things in machine learning.

Beta Distribution

- For Bernoulli likelihoods, most common prior is beta distribution:

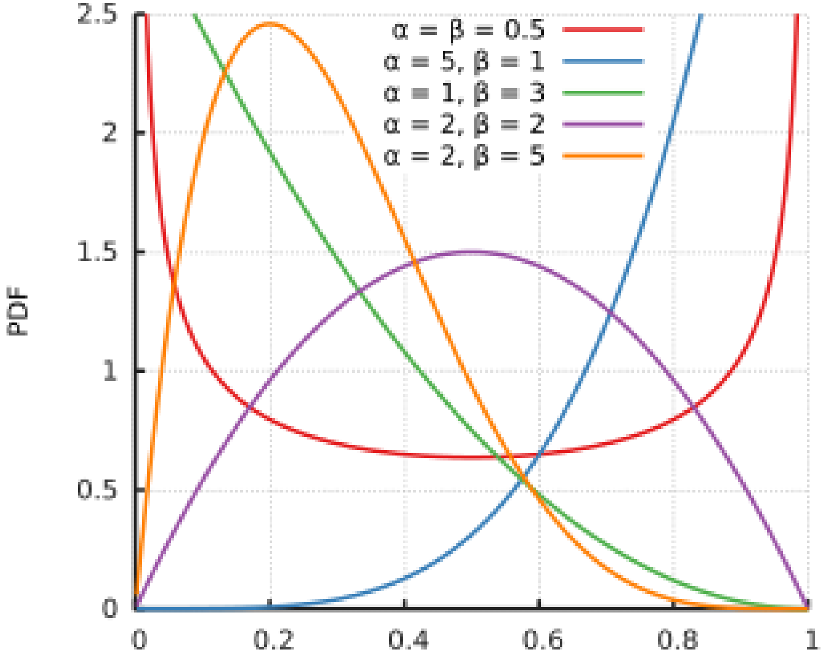
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } \underbrace{0 \leq \theta \leq 1}_{p(\theta \mid \alpha,\beta)=0 \text{ if } \theta<0 \text{ or } \theta>1}, \ \alpha>1, \ \beta>1$$

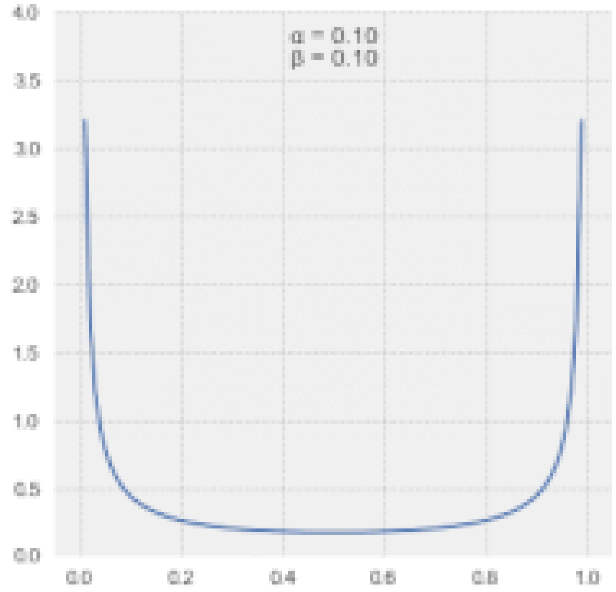
- Looks like a Bernoulli likelihood, with (α – 1) ones and (β-1) zeroes.
- Key difference with the Bernoulli is on the left side:
  - It defines a PDF over real numbers θ in the range 0 through 1.
    - Beta distribution is not assigning probabilities to binary values, but to PDF of θ.
      - "Probability over probabilities".
- From the "digression", we can resolve what is hidden in the ∝ sign:

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \leftarrow \text{"beta" function}$$

Beta Distribution

- The beta distribution for different choices of α and β:





- Why is using the beta distribution as prior so popular?
  - Fake reason: it is quite flexible, so can encode a variety of priors.
    - Can represent bias towards 0.5, towards 1 or 0, towards 0.2, towards only 1, or uniform if $\alpha = \beta = 1$.
    - But it is still limited. For example, you can't say that "the exact value 0.5 is particularly likely".

https://en.wikipedia.org/wiki/Beta_distribution

Posterior for Bernoulli Likelihood and Beta Prior

- Real reason people use the beta: posterior and MAP have simple forms.
  - The posterior with a Bernoulli likelihood and beta prior:

$$p(\theta \mid X, \alpha, \beta) \propto \underbrace{p(X \mid \theta)}_{} p(\theta \mid \alpha, \beta) \propto \theta^{n_1}(1-\theta)^{n_0}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

We assume that X is independent of α and β given θ

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

$$= \theta^{\tilde{\alpha}-1}(1-\theta)^{\tilde{\beta}-1}$$

  - This is another beta distribution with "updated" parameters $\tilde{\alpha}$ and $\tilde{\beta}$.
    - Where $\tilde{\alpha} = n_1 + \alpha$ and $\tilde{\beta} = n_0 + \beta$.
  - How do we know that this is a beta distribution?
    - Because constant in $\propto$ is unique.
      - "If you are proportional to a beta distribution, you are a beta distribution."
    - Make sure you understand why posterior is a beta distribution (important in this course).

Summary

- MAP Estimation:
  - Find parameters maximizing probability of parameters given data.
    - The “posterior”.
  - Requires prior distribution on parameters:
    - Can be used as bias towards parameters that overfit less.
- Probability review:
  - Product rule, marginalization rule, Bayes rule.
  - Continuous “probabilities” and how “$\propto$” has a restricted meaning for probabilities.
- Beta distribution:
  - Prior for Bernoulli that yields a closed-form posterior (another beta distribution).

- Next time: end the streak of “numbers of lectures with no MNIST digits”.