# CSC321 Lecture 18: Learning Probabilistic Models

Roger Grosse

# Overview

- So far in this course: mainly supervised learning
- Language modeling was our one unsupervised task; we broke it down into a series of prediction tasks
  - This was an example of distribution estimation: we'd like to learn a distribution which looks as much as possible like the input data.
- This lecture: basic concepts in probabilistic modeling
  - This will be review if you've taken 411.
- Following two lectures: more recent approaches to unsupervised learning

# Maximum Likelihood

- We already used maximum likelihood in this course for training language models. Let's cover it in a bit more generality.
- Motivating example: estimating the parameter of a biased coin
  - You flip a coin 100 times. It lands heads $N_H = 55$ times and tails $N_T = 45$ times.
  - What is the probability it will come up heads if we flip again?
- Model: flips are independent Bernoulli random variables with parameter $\theta$.
  - Assume the observations are independent and identically distributed (i.i.d.)

# Maximum Likelihood

- The likelihood function is the probability of the observed data, as a function of $\theta$.
- In our case, it's the probability of a *particular* sequence of H's and T's.
- Under the Bernoulli model with i.i.d. observations,

$$L(\theta) = p(\mathcal{D}) = \theta^{N_H}(1-\theta)^{N_T}$$

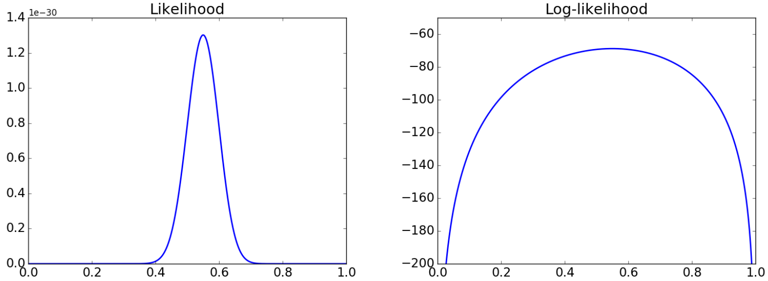
- This takes very small values (in this case, $L(0.5) = 0.5^{100} \approx 7.9 \times 10^{-31}$)
- Therefore, we usually work with log-likelihoods:

$$\ell(\theta) = \log L(\theta) = N_H \log \theta + N_T \log(1-\theta)$$

- Here, $\ell(0.5) = \log 0.5^{100} = 100 \log 0.5 = -69.31$

# Maximum Likelihood

$N_H = 55, \; N_T = 45$



# Maximum Likelihood

- Good values of $\theta$ should assign high probability to the observed data. This motivates the maximum likelihood criterion.
- Remember how we found the optimal solution to linear regression by setting derivatives to zero? We can do that again for the coin example.

$$\begin{aligned}\frac{\mathrm{d}\ell}{\mathrm{d}\theta} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(N_H \log\theta + N_T \log(1-\theta)\right) \\ &= \frac{N_H}{\theta} - \frac{N_T}{1-\theta}\end{aligned}$$

- Setting this to zero gives the maximum likelihood estimate:

$$\hat{\theta}_{\mathrm{ML}} = \frac{N_H}{N_H + N_T},$$

# Maximum Likelihood

- This is equivalent to minimizing cross-entropy. Let $t_i = 1$ for heads and $t_i = 0$ for tails.

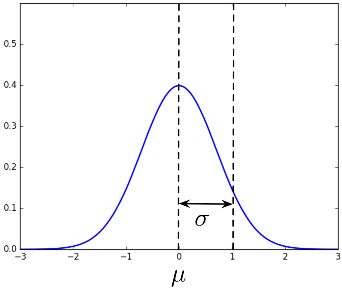
$$\begin{aligned}
\mathcal{L}_{CE} &= \sum_i -t_i \log \theta - (1 - t_i) \log(1 - \theta) \\
&= -N_H \log \theta - N_T \log(1 - \theta) \\
&= -\ell(\theta)
\end{aligned}$$

# Maximum Likelihood

- Recall the Gaussian, or normal, distribution:

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- The Central Limit Theorem says that sums of lots of independent random variables are approximately Gaussian.
- In machine learning, we use Gaussians a lot because they make the calculations easy.



# Maximum Likelihood

- Suppose we want to model the distribution of temperatures in Toronto in March, and we've recorded the following observations:

  -2.5 -9.9 -12.1 -8.9 -6.0 -4.8 2.4

- Assume they're drawn from a Gaussian distribution with known standard deviation $\sigma = 5$, and we want to find the mean $\mu$.
- Log-likelihood function:

$$\begin{aligned}\ell(\mu) &= \log \prod_{i=1}^{N} \left[ \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left( -\frac{(x^{(i)} - \mu)^2}{2\sigma^2} \right) \right] \\ &= \sum_{i=1}^{N} \log \left[ \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left( -\frac{(x^{(i)} - \mu)^2}{2\sigma^2} \right) \right] \\ &= \sum_{i=1}^{N} \underbrace{-\frac{1}{2} \log 2\pi - \log \sigma}_{\text{constant!}} - \frac{(x^{(i)} - \mu)^2}{2\sigma^2}\end{aligned}$$

# Maximum Likelihood

- Maximize the log-likelihood by setting the derivative to zero:

$$0 = \frac{\mathrm{d}\ell}{\mathrm{d}\mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \frac{\mathrm{d}}{\mathrm{d}\mu} (x^{(i)} - \mu)^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{N} x^{(i)} - \mu$$

- Solving we get $\mu = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$
- This is just the mean of the observed values, or the empirical mean.

# Maximum Likelihood

- In general, we don't know the true standard deviation $\sigma$, but we can solve for it as well.
- Set the *partial* derivatives to zero, just like in linear regression.

$$0 = \frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^{N} x^{(i)} - \mu$$

$$\begin{aligned}
0 = \frac{\partial \ell}{\partial \sigma} &= \frac{\partial}{\partial \sigma} \left[ \sum_{i=1}^{N} -\frac{1}{2} \log 2\pi - \log \sigma - \frac{1}{2\sigma^2} (x^{(i)} - \mu)^2 \right] \\
&= \sum_{i=1}^{N} -\frac{1}{2} \frac{\partial}{\partial \sigma} \log 2\pi - \frac{\partial}{\partial \sigma} \log \sigma - \frac{\partial}{\partial \sigma} \frac{1}{2\sigma} (x^{(i)} - \mu)^2 \\
&= \sum_{i=1}^{N} 0 - \frac{1}{\sigma} + \frac{1}{\sigma^3} (x^{(i)} - \mu)^2 \\
&= -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (x^{(i)} - \mu)^2
\end{aligned}$$

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

$$\hat{\sigma}_{\mathrm{ML}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \mu)^2}$$

# Maximum Likelihood

- So far, maximum likelihood has told us to use empirical counts or statistics:
  - **Bernoulli:** $\theta = \frac{N_H}{N_H + N_T}$
  - **Gaussian:** $\mu = \frac{1}{N} \sum x^{(i)}$, $\sigma^2 = \frac{1}{N} \sum (x^{(i)} - \mu)^2$
- This doesn't always happen; e.g. for the neural language model, there was no closed form, and we needed to use gradient descent.
- But these simple examples are still very useful for thinking about maximum likelihood.

## Data Sparsity

- Maximum likelihood has a pitfall: if you have too little data, it can overfit.
- E.g., what if you flip the coin twice and get H both times?

$$\theta_{\mathrm{ML}} = \frac{N_H}{N_H + N_T} = \frac{2}{2+0} = 1$$

- Because it never observed T, it assigns this outcome probability 0. This problem is known as data sparsity.
- If you observe a single T in the test set, the likelihood is $-\infty$.

# Bayesian Parameter Estimation

- In maximum likelihood, the observations are treated as random variables, but the parameters are not.
- The Bayesian approach treats the parameters as random variables as well.
- To define a Bayesian model, we need to specify two distributions:
  - The prior distribution $p(\boldsymbol{\theta})$, which encodes our beliefs about the parameters *before* we observe the data
  - The likelihood $p(\mathcal{D} \mid \boldsymbol{\theta})$, same as in maximum likelihood
- When we update our beliefs based on the observations, we compute the posterior distribution using Bayes' Rule:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta}') p(\mathcal{D} \mid \boldsymbol{\theta}') \, \mathrm{d}\boldsymbol{\theta}'}.$$

- We rarely ever compute the denominator explicitly.

# Bayesian Parameter Estimation

- Let's revisit the coin example. We already know the likelihood:

$$L(\theta) = p(\mathcal{D}) = \theta^{N_H}(1-\theta)^{N_T}$$

- It remains to specify the prior $p(\theta)$.
  - We can choose an uninformative prior, which assumes as little as possible. A reasonable choice is the uniform prior.
  - But our experience tells us 0.5 is more likely than 0.99. One particularly useful prior that lets us specify this is the beta distribution:

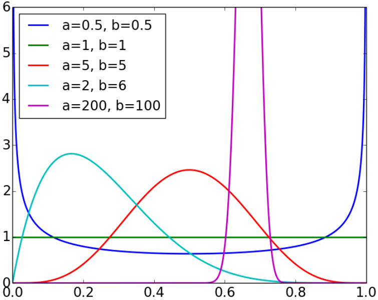
$$p(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}.$$

  - This notation for proportionality lets us ignore the normalization constant:

$$p(\theta; a, b) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

# Bayesian Parameter Estimation

- Beta distribution for various values of $a$, $b$:




- Some observations:
  - The expectation $\mathbb{E}[\theta] = a/(a+b)$.
  - The distribution gets more peaked when $a$ and $b$ are large.
  - The uniform distribution is the special case where $a = b = 1$.
- The main thing the beta distribution is used for is as a prior for the Bernoulli distribution.

# Bayesian Parameter Estimation

- Computing the posterior distribution:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathcal{D}) &\propto p(\boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{\theta}) \\ &\propto \left[\theta^{a-1}(1-\theta)^{b-1}\right] \left[\theta^{N_H}(1-\theta)^{N_T}\right] \\ &= \theta^{a-1+N_H}(1-\theta)^{b-1+N_T}. \end{aligned}$$

- This is just a beta distribution with parameters $N_H + a$ and $N_T + b$.
- The posterior expectation of $\theta$ is:

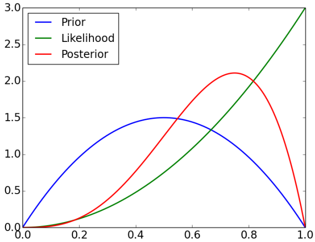
$$\mathbb{E}[\theta \mid \mathcal{D}] = \frac{N_H + a}{N_H + N_T + a + b}$$

- The parameters $a$ and $b$ of the prior can be thought of as pseudo-counts.
  - The reason this works is that the prior and likelihood have the same functional form. This phenomenon is known as conjugacy, and it's very useful.

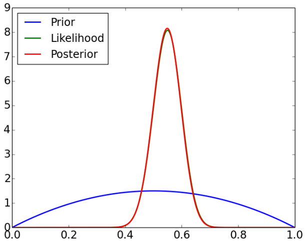
# Bayesian Parameter Estimation

Bayesian inference for the coin flip example:

Small data setting
$N_H = 2,\ N_T = 0$




Large data setting
$N_H = 55,\ N_T = 45$




When you have enough observations, the data overwhelm the prior.

# Bayesian Parameter Estimation

- What do we actually do with the posterior?
- The posterior predictive distribution is the distribution over future observables given the past observations. We compute this by marginalizing out the parameter(s):

$$p(\mathcal{D}' \,|\, \mathcal{D}) = \int p(\boldsymbol{\theta} \,|\, \mathcal{D}) p(\mathcal{D}' \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \tag{1}$$

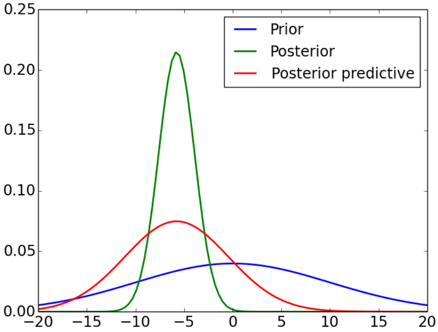
- For the coin flip example:

$$\begin{aligned} \theta_{\text{pred}} &= \Pr(x' = H \,|\, \mathcal{D}) \\ &= \int p(\theta \,|\, \mathcal{D}) \Pr(x' = H \,|\, \theta) \, \mathrm{d}\theta \\ &= \int \text{Beta}(\theta; N_H + a, N_T + b) \cdot \theta \, \mathrm{d}\theta \\ &= \mathbb{E}_{\text{Beta}(\theta; N_H + a, N_T + b)}[\theta] \\ &= \frac{N_H + a}{N_H + N_T + a + b}, \end{aligned} \tag{2}$$

# Bayesian Parameter Estimation

Bayesian estimation of the mean temperature in Toronto

- Assume observations are i.i.d. Gaussian with known standard deviation $\sigma$ and unknown mean $\mu$
- Broad Gaussian prior over $\mu$, centered at 0
- We can compute the posterior and posterior predictive distributions analytically (full derivation in notes)
- Why is the posterior predictive distribution more spread out than the posterior distribution?



# Bayesian Parameter Estimation

Comparison of maximum likelihood and Bayesian parameter estimation

- The Bayesian approach deals better with data sparsity
- Maximum likelihood is an optimization problem, while Bayesian parameter estimation is an integration problem
    - This means maximum likelihood is much easier in practice, since we can just do gradient descent
    - Automatic differentiation packages make it really easy to compute gradients
    - There aren't any comparable black-box tools for Bayesian parameter estimation (although Stan can do quite a lot)

# Maximum A-Posteriori Estimation

- Maximum a-posteriori (MAP) estimation: find the most likely parameter settings under the posterior
- This converts the Bayesian parameter estimation problem into a maximization problem

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{\theta} \,|\, \mathcal{D}) \\
&= \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{\theta}, \mathcal{D}) \\
&= \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{\theta}) \, p(\mathcal{D} \,|\, \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \; \log p(\boldsymbol{\theta}) + \log p(\mathcal{D} \,|\, \boldsymbol{\theta})
\end{aligned}$$

# Maximum A-Posteriori Estimation

- Joint probability in the coin flip example:

$$\begin{aligned}\log p(\theta, \mathcal{D}) &= \log p(\theta) + \log p(\mathcal{D} \,|\, \theta) \\ &= \text{const} + (a-1)\log\theta + (b-1)\log(1-\theta) + N_H \log\theta + N_T \log(1-\theta) \\ &= \text{const} + (N_H + a - 1)\log\theta + (N_T + b - 1)\log(1-\theta)\end{aligned}$$

- Maximize by finding a critical point

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta} \log p(\theta, \mathcal{D}) = \frac{N_H + a - 1}{\theta} - \frac{N_T + b - 1}{1 - \theta}$$

- Solving for $\theta$,

$$\hat{\theta}_{\mathrm{MAP}} = \frac{N_H + a - 1}{N_H + N_T + a + b - 2}$$

# Maximum A-Posteriori Estimation

Comparison of estimates in the coin flip example:

| | **Formula** | $N_H = 2, N_T = 0$ | $N_H = 55, N_T = 45$ |
|---|---|---|---|
| $\hat{\theta}_{\mathrm{ML}}$ | $\frac{N_H}{N_H+N_T}$ | $1$ | $\frac{55}{100} = 0.55$ |
| $\theta_{\mathrm{pred}}$ | $\frac{N_H+a}{N_H+N_T+a+b}$ | $\frac{4}{6} \approx 0.67$ | $\frac{57}{104} \approx 0.548$ |
| $\hat{\theta}_{\mathrm{MAP}}$ | $\frac{N_H+a-1}{N_H+N_T+a+b-2}$ | $\frac{3}{4} = 0.75$ | $\frac{56}{102} \approx 0.549$ |

$\hat{\theta}_{\mathrm{MAP}}$ assigns nonzero probabilities as long as $a, b > 1$.

# Maximum A-Posteriori Estimation

Comparison of predictions in the Toronto temperatures example