Lecture 8, Feb 13, 2024

Probability Density Estimation

- Previously we've considered learning problems using a loss function perspective; now we would like to consider a statistical perspective
- We begin by looking at density estimation problems
- Given a dataset $\mathcal{D} = \{\, \boldsymbol{x}^{(i)} \,\}_{i=1}^{N}$, we would like to determine the distribution generating this data
  - Assume θ is a hypothesis class that parametrizes the density function
  - $\mathcal{P}_{\boldsymbol{\theta}} = \{\, p(\boldsymbol{x} \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Gamma \,\}$

Maximum Likelihood Estimation

- In ML we aim to find the parameter value $\hat{\boldsymbol{\theta}}$ for which the observed data has the highest probability/density of occurring
- $\hat{\boldsymbol{\theta}}_{ML} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Gamma} p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)} | \boldsymbol{\theta})$
  - The term $p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)} | \boldsymbol{\theta})$ is known as the *likelihood function*
- We often assume that the data is *independently and identically distributed* (IID), which allows us to decompose the likelihood into a product
- Assuming IID, $p(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)} | \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)} | \boldsymbol{\theta})$
- Maximizing the likelihood is the same as maximizing the log of the likelihood function; this is referred to as *log-likelihood*
  - $\log p(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)} | \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left(p(\boldsymbol{x}^{(i)} | \boldsymbol{\theta}) \right)$
  - Practically, using log-likelihood prevents instability due to underflow (multiplying many very small numbers)
- To solve for the ML estimator we simply differentiate and set the derivative to zero
  - $\sum_{i=1}^{N} \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{x}^{(i)} | \boldsymbol{\theta})}{p(\boldsymbol{x}^{(i)} | \boldsymbol{\theta})} = 0$
  - In special cases we may obtain analytical solutions using linear algebra, but in general we may have to use nonlinear optimization methods
- MLE can also be used to perform regression
  - Consider observations as $y(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$ where $\epsilon \in \mathcal{N}(0, \sigma^2)$
  - $\hat{f}(\boldsymbol{x}, \boldsymbol{w})$ is the underlying function; we add some noise ϵ to get the measurement
  - $p(y | \boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(y | \hat{f}(\boldsymbol{x}, \boldsymbol{w}), \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \hat{f}(\boldsymbol{x}, \boldsymbol{w}))^2}{2\sigma^2} \right)$
  - The goal is to estimate the parameters $\boldsymbol{w}$
  - $$\begin{aligned} p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{w}, \sigma^2) &= \prod_{i=1}^{N} \mathcal{N}(y^{(i)} | \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}), \sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}))^2 \right) \end{aligned}$$
    - $\boldsymbol{y}$ is a column vector of all the observations while $\boldsymbol{X}$ has each of the $\boldsymbol{x}^{(i)}$ vectors as its rows
  - The negative log-likelihood is $\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}))^2 + N \log \sigma + \frac{N}{2} \log 2\pi$
    - Notice that the first term is just the l_2 loss function
    - The other two terms are constant in $\boldsymbol{w}$, so we see that MLE is equivalent to using a l_2 loss function when the data is IID Gaussian
  - This also lets us estimate the variance of the noise by differentiating the log-likelihood wrt σ^2 and

solve for zero

    * $-\frac{1}{2\sigma^3}\sum_{i=1}^{N}(y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}))^2 + \frac{N}{\sigma} = 0$
    * $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}))^2$

- For regression, we get a constant variance, so the error bars are constant size throughout the data
  - This is not reasonable since we expect the error bars to be smaller where we have more data points
  - Near the middle where we have more data, we should get smaller error while near the edges we should expect more
- Example exercise: assume IID Laplacian noise, formulate an optimization problem and solve for $\hat{\boldsymbol{w}}_{ML}$
  - The Laplace distribution is given by $\text{Lap}(\epsilon|\mu, b) = \frac{1}{2b}e^{-\frac{|\epsilon-\mu|}{b}}$
    * Mean, variance of $2b^2$
  - Get the joint likelihood: $$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, 2b^2) = \prod_{i=1}^{N}\text{Lap}(y^{(i)}|\hat{f}(\boldsymbol{x}, \boldsymbol{w}), 2b^2) = \left(\frac{1}{2b}\right)^N \exp\left(-\sum_{i=1}^{N}\frac{|y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w})|}{b}\right)$$
  - Negative log likelihood: $-\log(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, 2b^2)) = N\log 2b - \frac{1}{b}\sum_{i=1}^{N}|y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w})|$
  - Optimization problem: $$\hat{\boldsymbol{w}}_{ML} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, N\log 2b - \frac{1}{b}\sum_{i=1}^{N}|y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w})| = \underset{\boldsymbol{w}}{\operatorname{argmin}}\sum_{i=1}^{N}|y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w})|$$
    * Notice that this is akin to minimizing an l_1 loss function
    * This is no longer solvable analytically
- Example: Given measurements $x^{(1)} = 1, x^{(2)} = 2, x^{(3)} = 3, x^{(4)} = 3, x^{(5)} = 4$ distributed according to an exponential distribution $\rho e^{-\rho x}$, find the MLE of ρ
  - $p(x^{(1)}, \dots, x^{(5)}|\rho) = \prod_{i=1}^{5}\rho e^{-\rho x^{(i)}} = \rho^5 e^{-13\rho}$
  - NLL: $-\log(p(x^{(1)}, \dots, x^{(5)}|\rho)) = -5\log\rho + 13\rho$
  - Differentiate: $-\frac{5}{\rho} + 13 = 0 \implies \hat{\rho}_{\text{ML}} = \frac{5}{13}$

Maximum a Posteriori (MAP) Estimation

- In MAP estimation, we aim to find the parameter value that is most likely to occur given the data and a prior distribution of the parameter value
- $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$
  - The evidence/marginal likelihood in the denominator is often hard to compute
  - However for MAP we don't need to compute it since it does not depend on $\boldsymbol{\theta}$
- $\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
  - When the prior is uniform, this is equivalent to MLE
- Consider regression with a Gaussian prior and noise:
  - $p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha\boldsymbol{1}) = \left(\frac{1}{\sqrt{2\pi\alpha}}\right)^M \exp\left(-\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\alpha}\right)$
  - $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \prod_{i=1}^{N}\mathcal{N}(y^{(i)}|\hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}), \sigma^2)$

- The posterior is proportional to the product of the two
- $\hat{\boldsymbol{w}}_{MAP} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} (\hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}) - y^{(i)})^2 + \frac{1}{2\alpha} \boldsymbol{w}^T \boldsymbol{w}$

  $= \underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (\hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}) - y^{(i)})^2 + \frac{\sigma^2}{2\alpha} \boldsymbol{w}^T \boldsymbol{w}$
    * Notice that the first term is the l_2 loss function while the second is l_2 regularization
    * MAP estimation is equivalent to using an l_2 loss function with l_2 regularization, assuming a zero-mean Gaussian prior and IID Gaussian noise distribution
    * In the statistical perspective we are saying that we believe the weights are small prior to seeing the data; in the loss function perspective we are forcing the weights to be small

- Now consider a Laplace prior: $p(\boldsymbol{w}|\alpha) = \prod_{i=1}^{M} \text{Lap}(w_i|0, \alpha) = \left(\frac{1}{2\alpha}\right)^M \exp\left(-\frac{1}{\alpha}\sum_{i=1}^{M}|w_i|\right)$
  - Likelihood: $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y^{(i)}|\hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}), \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}}$
  - Negative log likelihood of posterior: $-\log(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2)) - \log(p(\boldsymbol{w}|\alpha))$
    * $\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^{(i)} - \hat{f}(\boldsymbol{x}^{(i)}, \boldsymbol{w}))^2 - \frac{1}{\alpha} \sum_{i=1}^{M} |w_i|$
  - We see again that this is equivalent to using l_2 loss with l_1 regularization with $\lambda = \frac{2\sigma^2}{\alpha}$

Frequentist vs. Bayesian Estimation

- In the frequentist approach, we assume that there exists a true fixed parameter value θ^*
  - We can get error bars by considering the distribution of possible datasets given this parameter value
  - However the error bars are not very good because they are independent of the inputs
  - Both MLE and MAP are frequentist methods since they give point estimates
- In the Bayesian approach, we use a single observation dataset to estimate the entire posterior distribution
  - This gives us both the mean as an estimate and a measure of uncertainty
  - Enables leveraging of priors