Gili Rusak, Alex Tsun
CS109

Section #8 Concept Check Solutions

1 Lecture 21, 2-26-20: Parameters and MLE

Suppose $x_1, \dots, x_n$ are iid samples from some distribution with density function $f_X(x;\theta)$, where θ is unknown. Recall that the likelihood of the data is

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)$$

Recall we solve an optimization problem to find $\hat{\theta}$ which maximizes L.

1. Write an expression for the log-likelihood, $LL(\theta) = \log L(\theta)$.
2. Why can we optimize $LL(\theta)$ rather than $L(\theta)$?
3. Why do we optimize $LL(\theta)$ rather than $L(\theta)$?

1. $LL(\theta) = \sum_{i=1}^{n} \log f_X(x_i;\theta)$
2. Logarithms are monotonic. This means that if $f(a) > f(b)$, then $\log(f(a)) > \log(f(b))$, so correctness of arg max is preserved.
3. Logs turn products into sums, which makes taking the derivative much simpler.

2 Lecture 22, 2-28-20: Gradient Ascent

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function which maps vectors $x \in \mathbb{R}^n$ to scalars $f(x) \in \mathbb{R}$.

1. What is the gradient ascent update step, with learning rate η?
2. Intuitively, what problem is gradient ascent trying to solve numerically?
3. What are some tradeoffs between a high and low learning rate (η)?

1. $x \leftarrow x + \eta \nabla f(x)$
2. We are attempting to numerically find the value of x that maximizes $f(x)$ by incrementally taking small steps in the direction of steepest ascent (according the the derivative).
3. A small learning rate might require more steps until convergence, while a large learning rate might overshoot and miss the absolute maximum.

3 Lecture 23, 3-2-20: Maximum A Posteriori

1. Intuitively, what is MAP? What problem is it trying to solve? How does it differ from MLE?

2. Given a 6-sided die (possibly unfair), you roll the die N times and observe the counts for each of the 6 outcomes as $n_1, ..., n_6$. What is the maximum a posteriori estimate of this distribution, using Laplace smoothing? Recall that the die rolls themselves follow a multinomial distribution.

1. From the course notes: The paradigm of MAP is that we should choose the value for our parameters that is **the most likely given the data**. At first blush this might seem the same as MLE; however, remember that MLE chooses the value of parameters that **makes the data most likely**. One of the disadvantages of MLE is that it best explains data we have seen and makes no attempt to generalize to unseen data. In MAP, we incorporate prior belief about our parameters, and then we update our posterior belief of the parameters based on the data we have seen.

2. Using a prior which represents one imagined observation of each outcome is called "Laplace smoothing" and it guarantees that none of your probabilities are 0 or 1. The Laplace estimate for a Multinomial RV is $p_i = \frac{n_i+1}{N+6}$ for $i = 1, ..., 6$.

4 Lecture 24, 3-4-20: Naive Bayes

Recall the classification setting: we have data vectors of the form $X = (X_1, \ldots, X_d)$ and we want to predict a label $Y \in \{0, 1\}$.

1. Recall in Naive Bayes, given a data point x, we compute $P(Y = 1|X = x)$ and predict $Y = 1$ provided this quantity is ≥ 0.5, and otherwise we predict $Y = 0$. Decompose $P(Y = 1|X = x)$ into smaller terms, and state where the Naive Bayes assumption is used.

2. Suppose we are given example vectors with labels provided. Give a formula to estimate (using maximum likelihood) each quantity $P(X_i = x_i|Y = y)$ above, for $i \in \{1, \ldots, d\}$ and $y \in \{0, 1\}$. You can assume there is a function count which takes in any number of boolean conditions and returns a count over the data of the number of examples in which they are true. For example, count$(X_3 = 2, X_5 = 7)$ returns the number of examples where $X_3 = 2$ and $X_5 = 7$.

1.

$$P(Y = 1|X = x) = \frac{P(Y = 1)P(X = x|Y = 1)}{P(Y = 1)P(X = x|Y = 1) + P(Y = 0)P(X = x|Y = 0)} \quad \text{(Bayes+LTP)}$$

$$= \frac{P(Y = 1)\prod_{i=1}^{d} P(X_i = x_i|Y = 1)}{P(Y = 1)\prod_{i=1}^{d} P(X_i = x_i|Y = 1) + P(Y = 0)\prod_{i=1}^{d} P(X_i = x_i|Y = 0)} \quad \text{(NB Assumption)}$$

2. $P(X_i = x_i|Y = y) = \dfrac{\text{count}(X_i = x_i, Y = y)}{\text{count}(Y = y)}$