

Machine Learning Worksheet 1

Probability Theory

1 Refresher

Lots of questions regarding uncertainty in machine learning can be answered by the two elementary rules of probability, known as the *sum rule* and the *product rule*.

$$\text{sum rule} \quad p(X = x) = \sum_y p(X = x, Y = y)$$

$$\text{product rule} \quad p(X = x, y = Y) = p(Y = y|X = x)p(X = x)$$

(X and Y are discrete random variables). From the product rule, one immediately obtains the so called *Bayes' rule*:

$$p(Y = y|X = x) = \frac{p(X = x|Y = y)p(Y = y)}{p(X = x)}$$

You also should be familiar with the *expectation* $\mathbb{E}[X]$ of a random variable X and its *variance* $\text{var}[X]$:

$$\mathbb{E}[X] = \sum_x xp(X = x)$$

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Note that we usually write $p(x)$ instead of $p(X = x)$.

In the case of continuous random variables, the summation in the above formulae is substituted by an integration with respect to the corresponding probability density.

2 Basic Probability

Problem 1. There are eleven urns labeled by $u \in \{0, 1, 2, \dots, 10\}$, each containing ten balls. Urn u contains u black balls and $10 - u$ white balls. Alice selects an urn u at random and draws N times with replacement from that urn, obtaining n_B black balls and $N - n_B$ white balls. If after $N = 10$ draws $n_B = 3$ black balls have been drawn, what is the probability that the urn Alice is using is urn u ?

Now, let Alice draw another ball from the same urn. What is the probability that the next drawn ball is black (show your work)?

Problem 2. Consider two variables X and Y with joint distribution $p(x, y)$. Prove the following two results:

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]] \tag{1}$$

$$\text{var}[X] = \mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]] \tag{2}$$

Here $\mathbb{E}_X[X|Y]$ denotes the expectation of X under the conditional distribution $p(x|y)$, with a similar notation for the conditional variance.

3 Probability Inequalities

Inequalities are useful for bounding quantities that might otherwise be hard to compute. We'll begin with a simple inequality, called the Markov inequality after Andrei A. Markov, a student of Pafnuty Chebyshev.

3.1 Markov Inequality

Let X be a non-negative, discrete random variable, and let $c > 0$ be a positive constant.

Problem 3. Show that

$$P(X > c) \leq \frac{\mathbb{E}[X]}{c}.$$

Now, consider flipping a fair coin n times. Using the Markov Inequality, what is the probability of getting more than $3/4n$ heads?

3.2 Chebyshev Inequality

Apply the Markov Inequality to the deviation of a random variable from its mean, i.e. for a general random variable X we wish to bound the probability of the event $\{|X - \mathbb{E}[X]| > a\}$, which is the same as the event $\{(X - \mathbb{E}[X])^2 > a^2\}$.

Problem 4. Prove that

$$P(|X - \mathbb{E}[X]| > a) \leq \frac{\text{var}(X)}{a^2}$$

holds. Again, consider flipping a fair coin n times. Now use the Chebyshev Inequality to bound the probability of getting more than $3/4n$ heads.

3.3 Jensen's Inequality

Let f be a convex function. If $\lambda_1, \dots, \lambda_n$ are positive numbers with $\lambda_1 + \dots + \lambda_n = 1$, then for any $x_1, \dots, x_n \in I$:

$$f(\lambda_1 x_1 + \dots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \dots + \lambda_n f(x_n)$$

Problem 5. Prove this inequality by using induction on n .

Problem 6. Using Jensen's Inequality, show that for a finite random variable X (with n different values), its entropy is always bounded above by $\ln n$. Additionally, prove that the Kullback-Leibler divergence between any two discrete probability distributions is always non-negative.