Statistics for Machine Learning (Draft)

1 Probability (Background)

A **random variable** is a variable whose value depends on the outcome of a random event, called an **experiment**. Mathematically, its domain is the **sample space**—the space of all possible outcomes—while its co-domain is (typically) the set of real numbers.

An example of an experiment is two fair coin flips, where the outcomes are $\{HH, HT, TH, TT\}$. An example of an associated random variable, call it W, is "the number of heads." The range of W is $\{0, 1, 2\}$. This random variable is discrete because the sample space is discrete.

A **probability distribution** is a mathematical function associated with a random variable whose values are probabilities over subsets of a sample space, called an **event**. More specifically, these probabilities are associated with the inverse of W: $W^{-1}(0) = \{TT\}$, $W^{-1}(1) = \{HT, TH\}$, $W^{-1}(2) = \{HH\}$. Since the coin is fair, $P(W = 0) = 1/4$, $P(W = 1) = 1/2$, and $P(W = 2) = 1/4$. We say that the random variable W is distributed according to P, and we write $W \sim P$.

As the name suggests, random variables vary! Indeed, you may observe different results if you run the same experiment multiple times. Still, what might you *expect* the value of a random variable to be? This quantity is called its **expectation**. And how might you expect the result of any one experiment to vary from this expectation. This quantity is called its **variance**.

The **expectation**, or **expected value**, of a discrete random variable with range $X = \{x_1, \ldots, x_n\}$ and probabilities $P = \{p_1, \ldots, p_n\}$ is computed as follows:

$$\mathbb{E}_{X \sim P}[X] = \sum_{x_i \in X} p_i x_i$$

For example, the $\mathbb{E}[W] = 1/4(0) + 1/2(1) + 1/4(2) = 1$. When the probability distribution is clear from context, we drop the subscript $X \sim P$. The expected value of a random variable is also called the **mean**.

The **variance** of a random variable is defined as the expected value of the squared difference between the random variable and its mean: $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$. By linearity of expectations—the expected value of the sum of random variables is equal to the sum of their expectations—this definition simplifies as follows:

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] && (1)\\
&= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])] && (2)\\
&= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] && (3)\\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 && (4)\\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && (5)
\end{aligned}$$

For example, since $(\mathbb{E}[W])^2 = 1$ and $\mathbb{E}[W^2] = 1/4(0^2) + 1/2(1^2) + 1/4(2^2) = 3/2$, $\text{Var}[W] = 1/2$.

Exercise Calculate the variance of a single flip of a fair coin, where the random variable of interest is again the number of heads.

Solution The expected number of heads in a fair coin flip is $1/2(0) + 1/2(1) = 1/2$. The variance is therefore $1/2(0^2) + 1/2(1^2) - (1/2)^2 = 1/4$.

Two common examples of discrete probability distributions are the Bernoulli distribution and the binomial distribution. A **Bernoulli trial** is an experiment with exactly two possible outcomes: success/failure, yes/no, true/false, 1/0, healthy/sick, etc. A Bernoulli random variable is distributed according to a **Bernoulli distribution**, where the probability of success is a constant p, and the probability of failure is $1 - p$.

The expected value of a Bernoulli random variable B is $\mathbb{E}[B] = p(1) + (1 - p)0 = p$. The variance of B is $\text{Var}[B] = \mathbb{E}[B^2] - (\mathbb{E}[B])^2 = p(1^2) + (1 - p)0^2 - p^2 = p - p^2 = p(1 - p)$.

The random variable W summarizes the outcome of not just one, but two experiments. Each experiment can be described by its own random variable, and importantly, these random variables are *independent and identically distributed (i.i.d.)*. A collection of random variables is i.i.d. if each is distributed according to the same probability distribution and each is independent of all the others, meaning the value of one does not influence the value of another.

A **binomial** random variable describes the result of n i.i.d. Bernoulli trials. The range of a binomial random variable is the number of successes. For example, W is a binomial random variable, where $n = 2$, and heads corresponds to success so that the range of outcomes is, once again, $\{0, 1, 2\}$.

The expected value of a binomial random variable X is the expected value of the n independent Bernoulli random variables: i.e., $\mathbb{E}[X] = \mathbb{E}[X_1 + \ldots + X_n] = n\mathbb{E}[X_1] = np$.

Fact The variance of the sum of *independent* random variables equals the sum of their variances. Hence, $\text{Var}[X] = \text{Var}[X_1 + \ldots + X_n] = \text{Var}[X_1] + \ldots + \text{Var}[X_n] = n\text{Var}[X_1] = np(1 - p)$.

When there are more than two possible outcomes, e.g., the roll of a k-sided die, the Bernoulli distribution is generalized by the **categorical** distribution, and the binomial, by the **multinomial**.

2 Supervised Learning, Revisited

In supervised learning, we are given a data set, $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^c, i \in \{1, \ldots, m\}\}$, based on which we build a machine learning model f intended to predict a value $y = f(x) \in \mathbb{R}^c$, given a value $x \in \mathbb{R}^d$. This model may be built heuristically, like a decision tree, or it may be based on a probabilistic model, like Naive Bayes. The goal, when building such a model, is typically to minimize some loss function. A typical loss function for a binary classifier is 0-1, or misclassification, loss:

$$\mathcal{L}(y, f(x)) = \begin{cases} 0 & y = f(x) \\ 1 & y \neq f(x) \end{cases}$$

More precisely, the goal when building a machine learning model is to minimize **risk**, meaning the *expected* loss—computed with respect to some underlying probability distribution from which the data were sampled. In other words, in supervised learning there always is an underlying assumption that the data set was generated according to some data-generating process. Moreover, there is a further assumption that the as-of-yet unseen data to which the model will be applied is drawn from an *identical* distribution. This assumption gives us some faith that a machine learning model built from existing data—m samples from this distribution—can **generalize** well, meaning perform well on as-of-yet unseen data. Without it, there would be no reason to believe that a machine learning model would serve any purpose at all.

One approach to minimizing risk is to minimize **empirical risk**, which is the risk on the data set $\mathcal{D}$:

$$\sum_{i=1}^{m} \mathcal{L}(y_i, f(x_i))$$

However, minimizing empirical risk does not ensure generalization, because a machine learning model can **overfit** the training data. That is, it might memorize idiosyncratic aspects of $\mathcal{D}$, rather than learning general rules that would work well on as-of-yet unseen data. For this reason, $\mathcal{D}$ is often split up into a training set and a test set; for example, 80% of the data might be allocated to training and the remaining 20%, to testing. Then various models can be trained on the training set, and evaluated on the test set, as a proxy for generalization error. This process is called **model selection**.

The further assumption that the data set $\mathcal{D}$ comprises *independent* samples justifies a further model selection technique called k**-fold cross-validation**. This technique involves partitioning the data not jut once, but k times, into a training set (e.g., 80% of the data) and a test set (e.g., 20% of the data). Each of these partitions is called a **fold**. One way to generate k folds might be to shuffle $\mathcal{D}$ k times before partitioning. Using k folds instead of just one is sensible only under the assumption that the data are i.i.d.. Otherwise, an arbitrary partitioning of $\mathcal{D}$ into k folds could obscure regularities in the data.

Model selection can not only to avoid overfitting; it can avoid underfitting as well. A model overfits a data set if it is too flexible, while it **underfits** a data set if it is so inflexible that it fails to identify relevant trends.

3 Statistical Modelling

Statistical machine learning methods like Naive Bayes rely on further statistical modelling assumptions, beyond just i.i.d.. A **statistical model** is a set of assumptions describing a data-generation process as a relationship among random variables. Mathematically, a statistical model is a sample space together with a *set*, or **family**, of probability distributions. A **parametric model**, or a **finite-dimensional model**, is a parameterized statistical model with a finite number of parameters.[1] Statistical estimation (and statistical machine learning) is concerned with estimating the parameters of statistical models from samples (i.e., data).

There will be a presidential election next month. Assume some proportion, say p, of voters plans to vote Democratic, while the remaining proportion, $1 - p$, of voters plans to vote Republican. A poll might be conducted to estimate this proportion, in order to predict the outcome of the election. This poll would randomly sample a subset of the population about their voting plans. We can model the outcome of this poll as a binomial random variable, meaning n i.i.d. Bernoulli trials, each with parameter p.[2] We can then use the poll data (i.e., the random samples) to estimate this parameter p.

A **statistic** is a quantity computed from data. An **estimator** is a rule for calculating a statistic, which in this context is called an **estimate**. Given a data set $\mathcal{D} = \{x_1, \ldots, x_n\}$ comprising the outcome of n Bernoulli trials (i.e., $x_i \in \{0, 1\}$, for all $i \in \{1, \ldots, n\}$), the **sample proportion** $\bar{x} = {}^1\!/_n \sum_{i=1}^{n} x_i$ is an example of an estimate, computed by the **sample proportion estimator**, a function which takes as input a data set of 0's and 1s and outputs its sample proportion.

Remark A Bernoulli random variable with parameter p is a parametric model, as it is a statistical model with one parameter. While model parameters, such as the proportion of Democratic voters, are often unknown, they are fixed, *not* random, quantities. Statistics, on the other hand, which depend on random samples, are random variables.

Next, we explore some of what makes for a good estimator. We do so in the context of the polling example, where sample proportion is a potential estimate of the success probability p of a binomial random variable.

[1] Infinite-dimensional statistical models are called **non-parametric models**.

[2] Although a mathematical ideal, in practice, no one knows how to sample i.i.d. from a population.

4 Maximum Likelihood Estimation

One way of estimating the parameters of a statistical model is via **maximum likelihood estimation (MLE)**. The idea of this approach is to find a parameter that maximizes the "likelihood" of the given data: i.e., find θ s.t. $\mathcal{L}_{\mathcal{D}}(\theta)$ is maximized: i.e.,

$$\theta^* \in \arg\max_{\theta \in \Theta} \mathcal{L}_{\mathcal{D}}(\theta)$$

Since log is a monotonic function, we can equivalently maximize the log likelihood:

$$\theta^* \in \arg\max_{\theta \in \Theta} \log \mathcal{L}_{\mathcal{D}}(\theta)$$

By the i.i.d. assumption, the likelihood simplifies as $\mathcal{L}_{\mathcal{D}}(\theta) = \prod_{i=1}^{m} \mathcal{L}_{\mathcal{D}_i}(\theta)$. For example, the likelihood function for a binomial random variable simplifies as the product of the likelihood functions of n Bernoullis. Ultimately, the MLE objective is:

$$\begin{aligned}\theta^* &\in \arg\max_{\theta \in \Theta} \log \prod_{i=1}^{m} \mathcal{L}_{(x_i, y_i)}(\theta) \\ &\in \arg\max_{\theta \in \Theta} \sum_{i=1}^{m} \log \mathcal{L}_{(x_i, y_i)}(\theta)\end{aligned}$$

But what, pray tell, is a likelihood function? To define a likelihood function, we assume an underlying statistical model, i.e., a family $\mathcal{P}_\theta(\mathcal{D})$ of probability distributions.

In the polling example, the underlying statistical model is a Bernoulli random variable X with parameter $p \in [0, 1]$, i.e., $P(X = 1) = p$ and $P(X = 0) = 1 - p$, which we write as the family of probability distributions:

$$\mathcal{P}_p(X) \doteq \begin{cases} p & \text{if } X = 1 \\ 1 - p & \text{if } X = 0 \end{cases}$$

We can rearrange this function as follows:

$$\mathcal{P}_p(X) = p^X (1 - p)^{1 - X}$$

As above, by case analysis, this reformulation evaluates to p when $X = 1$, and $1 - p$ when $X = 0$.

The likelihood function is now essentially the same as this statistical model, except that rather than fixing the parameter, and making the function dependent on the random variable, instead, the data (e.g., $x \in \{0, 1\}$) are assumed to be fixed/given, and the parameter is the input to the function: $\mathcal{L}_x(p) = p^x (1 - p)^{1 - x}$.

Therefore, assuming $x_1, \ldots, x_n$ are the outcomes of n i.i.d. Bernoulli trials,

$$\begin{aligned}\log \mathcal{L}_{\{x_i\}_{i=1}^n}(p) &= \log \prod_{i=1}^{n} \mathcal{L}_{x_i}(p) \\ &= \sum_{i=1}^{n} \log \mathcal{L}_{x_i}(p) \\ &= \sum_{i=1}^{n} \log \left\{ p^{x_i} (1 - p)^{1 - x_i} \right\} \\ &= \sum_{i=1}^{n} \left(x_i \log p + (1 - x_i) \log(1 - p) \right) \\ &= n \bar{x} \log p + n(1 - \bar{x}) \log(1 - p)\end{aligned}$$

Here, $\bar{x} = {}^1\!/_n \sum_{i=1}^n x_i$ is the **sample mean**, meaning an average computed from data (i.e., a sample), which in the case of a binomial random variable is called the **sample proportion**.

Next, let's optimize this log likelihood, by taking its derivative and setting it equal to zero.

The derivative is:

$$\begin{aligned} \frac{\partial \log \mathcal{L}_{\{x_i\}_{i=1}^n}(p)}{\partial p} &= \frac{\partial \left\{ n\bar{x} \log p + n(1-\bar{x}) \log(1-p) \right\}}{\partial p} \\ &= \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} \end{aligned}$$

Setting this derivative equal to zero yields:

$$\frac{n\bar{x}}{p^*} = \frac{n(1-\bar{x})}{1-p^*}$$

Finally, $\bar{x}(1-p^*) = p^*(1-\bar{x})$, so $p^* = \bar{x}$.

As our intuition suggests, the MLE estimator of the parameter p of a binomial random variable is the sample proportion, i.e., the number of successes divided by the number of trials.