第八讲
概率论（II）
Lecture 8
Probability (II)

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

- 本讲义在准备过程中由于时间所限，所用材料来源并未规范标示引用来源。所引材料仅用于教学所用，作者无意侵犯原著者之知识版权，所引材料之知识版权均归原著者所有；若原著者介意之，请联系作者更正及删除。

  The time limit during the preparation of these slides incurs the situation that not all the sources of the used materials (texts or images) are properly referenced or clearly manifested. However, all materials in these slides are solely for teaching and the author is with no intention to infringe the copyright bestowed on the original authors or manufacturers. All credits go to corresponding IP holders. Please address the author for any concern for remedy including deletion.

期望
Expectation

- 假设X是一个离散随机变量，取值为x_k的概率为$P(x_k)$，（$k=1,2,\cdots,n$）。X的期望值被定义为：

  Let X be a discrete random variable taking values x_k with probability $P(x_k)$, $(k=1,2,\cdots,n)$. The expected value of X is defined as:

$$E(X)=\sum_{i=1}^{n}x_kP(x_k)$$

- 如果X是一个连续随机变量，其概率密度函数为$f(x)$，那么期望值被定义为：

  If X is a continuous random variable with p.d.f. $f(x)$ the expected value is defined as:

$$E(X)=\int_{x=-\infty}^{+\infty}xf(x)dx$$

- 期望是一个线性算子，也就是说，对于任意m个随机变量的线性组合$\sum_{j=1}^{m}a_jX_j$，其期望值为：

  The expect value is a linear operator, that is, for any linear combination of m random variables:

$$E\left(\sum_{j=1}^{m}a_jX_j\right)=\sum_{j=1}^{m}a_jE(X_j)$$

方差
Expectation

- 离散随机变量X取值为x_k，概率为$P(x_k)$（$k=1,2,\cdots,n$）的方差被定义为：

  The variance of a discrete random variable X taking values x_k with probability $P(x_k)$ $(k=1,2,\cdots,n)$ is defined as:

$$V(X)=\sum_{i=1}^{n}\left(x_k-E(X)\right)^2 P(X=x_k)$$

- 如果X是一个连续随机变量，其概率密度函数为$f(x)$，那么方差被定义为：

  If X is a continuous random variable with p.d.f. $f(x)$, then:

$$V(X)=\int_{x=-\infty}^{+\infty}\left(x-E(X)\right)^2 f(x)dx$$

- 由于$\left(x_k-E(X)\right)^2$衡量了x_k与期望值$E(X)$之间的距离，方差是变量实际取值与期望值之间距离的加权平均，权重为概率。标准差$\sigma(X)$然后是变量X的值与期望值之间的欧氏距离。请注意，标准差的表示量纲与变量X的测量量纲相同。

  Since $\left(x_k-E(X)\right)^2$ measure the distance of x_k from the expected value $E(X)$, the variance is the average of the distances between the values that the variable can actually take and the expectation, weighted by the probabilities. The standard deviation $\sigma(X)$ is then the Euclidean distance between the values of the variable X and the expectation. Note that the standard deviation is expressed in the same measurement scale of the variable X.

性质
Properties

- 如果X服从参数为μ和σ^2的正态分布，那么可以证明$E(X)=\mu$，$V(X)=\sigma^2$。如果$X\sim\mathrm{N}(\mu,\sigma^2)$，且$Y=aX+b$，其中$a$和$b$为实数，则$Y\sim\mathrm{N}(a\mu+b,a^2\sigma^2)$。这个性质允许我们从标准正态分布$X\sim\mathrm{N}(0,1)$来生成任意正态分布。我们可以简单定义$Y=\mu+\sigma X$来生成满足$Y\sim\mathrm{N}(\mu,\sigma^2)$的变量$Y$。

  If X has a Normal distribution with parameters μ and σ^2, then it can be shown that $E(X)=\mu$ and $V(X)=\sigma^2$. If $X\sim\mathrm{N}(\mu,\sigma^2)$ and $Y=aX+b$ for real numbers a and b, then $Y\sim\mathrm{N}(a\mu+b,a^2\sigma^2)$. This property allows to generate any Normal distribution from the Standard Normal distribution $X\sim\mathrm{N}(0,1)$. For example, we can simply define a variable $Y=\mu+\sigma X$ to generate $Y\sim\mathrm{N}(\mu,\sigma^2)$.

- 假设X是一个随机向量，$X=(X_1,X_2,\cdots,X_C)$，其每个分量都是一个单随机变量。我们定义随机向量X的期望为$E(X)=\big(E(X_1),E(X_2),\cdots,E(X_C)\big)$。其中，$E(X_i)$的计算都是使用随机变量$X_i$的边缘概率分布函数进行的。

  Suppose X is a random vector, namely, $X=(X_1,X_2,\cdots,X_C)$, with each component a univariate random variable. Then we can define its expectation by the means of its components: $E(X)=\big(E(X_1),E(X_2),\cdots,E(X_C)\big)$. The calculation of $E(X_i)$'s are carried out using the marginal p.d.f. of X_i's:

$$E(X_i)=\begin{cases}\sum_j x_{i,j}P\left(X_i=x_{i,j}\right) & discrete\ R.V.\\ \int_{-\infty}^{+\infty} x_i f(x_i)dx_i & continuous\ R.V.\end{cases}$$

协方差
Covariance

- 随机变量X的方差是衡量X围绕$E(X)$的变动程度的度量。现在假设我们有两个离散随机变量X_1和X_2，如果我们想考查X_1与X_2的变动是否一致，则我们可以考虑具有联合概率分布函数$P(X_1, X_2)$的向量$X = (X_1, X_2)$，通过定义协方差，作为它们联合变动性的指标。

  The variance of a random variable X is a measure of the variability of X around $E(X)$. Suppose now we have two discrete random variables X_1 and X_2. If we want to measure whether the directions of variabilities of X_1 and X_2 are coincident or not, we can do it by considering the vector $X = (X_1, X_2)$ with joint probability distribution $P(X_1, X_2)$. By defining covariance we have a index to measure their joint variability.

$$\operatorname{cov}(X_1, X_2) = \begin{cases} \sum_{i,j} \left(x_{1,i} - E(X_1)\right)\left(x_{2,j} - E(X_2)\right) P\left(x_{1,i}, x_{2,j}\right) & For\ discrete\ R.V. \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x_1 - E(X_1))(x_2 - E(X_2)) f(x_1, x_2) dx_1 dx_2 & For\ continuous\ R.V. \end{cases}$$

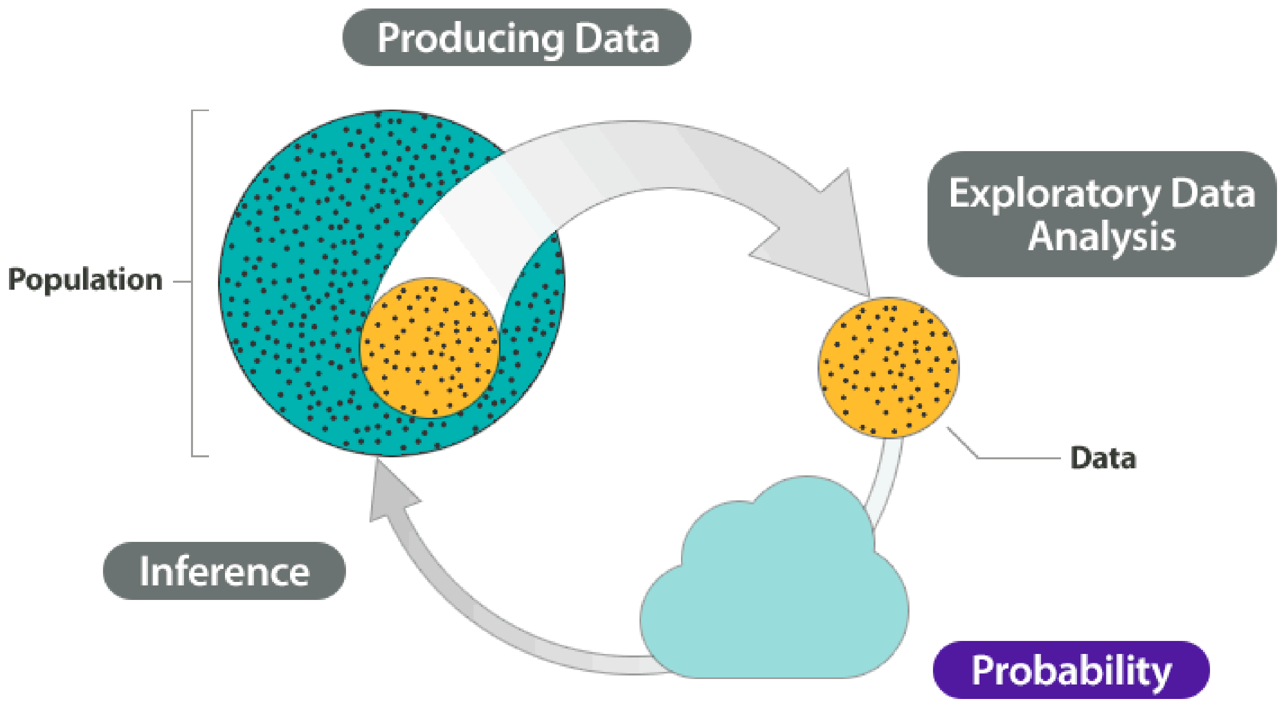
- 实际采样计算时，我们一般采用下面公式：

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

概率
Probability

- 由于概率分布函数或概率密度函数刻画了概率模型的行为，因此在实际问题中我们往往需要求出这些函数的表达。下图展示了实际中如何求得概率分布。

  Since the probability distribution function or probability density function depicts the dynamics and properties of the stochastic model, it is vital to find or estimate these functions in practice. The diagram below illustrates how to derive probability distribution in practice.



贝叶斯决策理论
Bayesian Decision Theory

- 贝叶斯决策理论是一种在不确定性下做出决策的方法。它结合了贝叶斯概率（使用新信息更新信念）和决策理论（根据结果和效用选择行动），通过考虑概率和潜在收益来帮助我们做出最优选择。贝叶斯决策理论是在考虑了不确定性和可用信息，用于找到最佳决策，最大化预期效用的一个框架。

  Bayesian decision theory is a method for making decisions under uncertainty. It combines Bayesian probability (updating beliefs with new information) and decision theory (choosing actions based on outcomes and utilities) to make optimal choices by considering probabilities and potential benefits. It's a framework used to find the best decision that maximizes expected utility, accounting for uncertainty and available information.

- 例子：如何判定鲈鱼与三文鱼。一捕鱼公司从近海捕捞一批鱼类，在产品线上做分拣。如何自动判定从传送出口送上来的是鲈鱼，还是三文鱼？

  How to classify see bass and salmon. A fishing company wants to sort the fishes caught in the near see on the assembly. How to decide a fish delivered from the conveyer is see bass or salmon?

贝叶斯决策理论
Bayesian Decision Theory

- 如果没有任何信息，则随机猜测，如果有对鲈鱼或三文鱼的先验认识，则根据这个这个认识做判断。例如，令X表示为何种鱼的随机变量，$X=0$表示鲈鱼，$X=1$表示三文鱼，如果根据先验知识，有$P(X=0)=0.25$，$P(X=1)=0.75$，则无论传送带传送上来什么鱼，统一猜三文鱼。为什么？

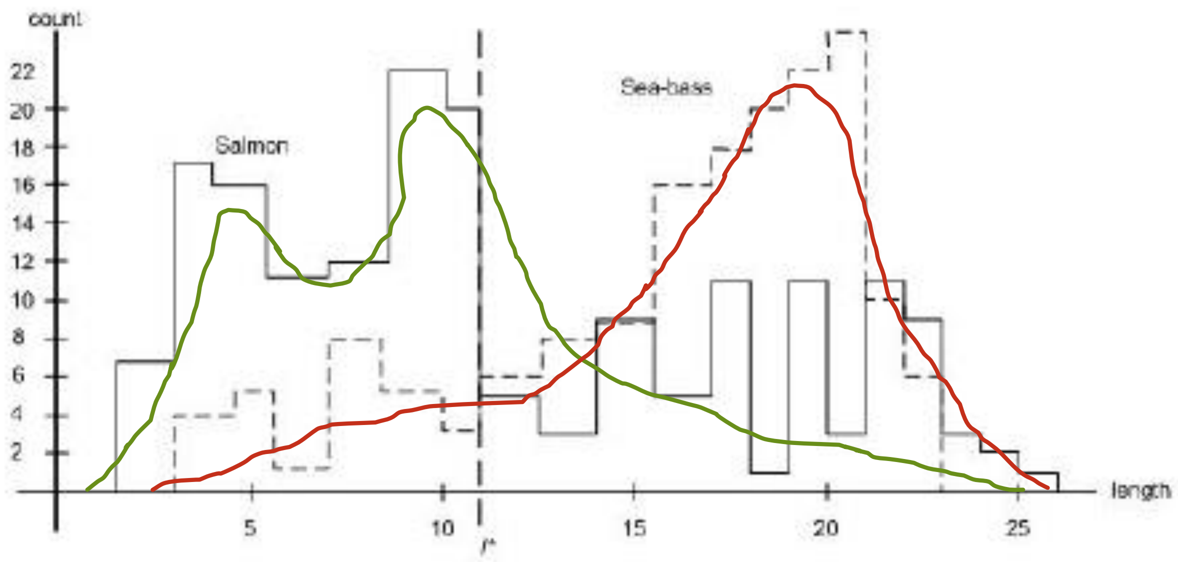
  If there is no any other information, the only way to perform is a random guess. But, if there exists some priori knowledge about see bass and salmon, the decision can take this priori into consideration. For example, let X be the random variable representing the kind of fish, and $X=0$ indicates the fish currently conveyed is see bass, while $X=0$ indicates salmon. If according to experience or survey, we have $P(X=0)=0.25$, $P(X=1)=0.75$, then a universal strategy is to always guess salmon. Why?

(a) (b)

贝叶斯决策理论
Bayesian Decision Theory

- 假设在长期的实践中，观察到鱼体长度与种类的关系。经平滑，得到以下概率分布函数。则又该如何判断呢？

  Suppose during the practice, the relation between the class of fish and its corresponding length has been drawn. The probability distributions of the length of fishes are shown below after smoothing. Then how to decide?



贝叶斯决策理论
Bayesian Decision Theory

- 我们已经学过条件概率，如概率$P(B|A)$描述的是事件B在另一个事件A已经发生的条件下的概率，记为$P(B|A) = P(AB)/P(A)$。A和B可能是相互独立的两个事件（即A和B的发生互不相关，此时$P(AB) = P(A)P(B)$），也可能不是。

  We have learned the conditional probability previously. For example, $P(B|A)$ depicts the probability of event B occurs given the happening of event A, denoted as $P(B|A) = P(AB)/P(A)$. A and B can be two independent events (aka, occurrence of event A impact nothing on event B and vice versa, now $P(AB) = P(A)P(B)$), or not.

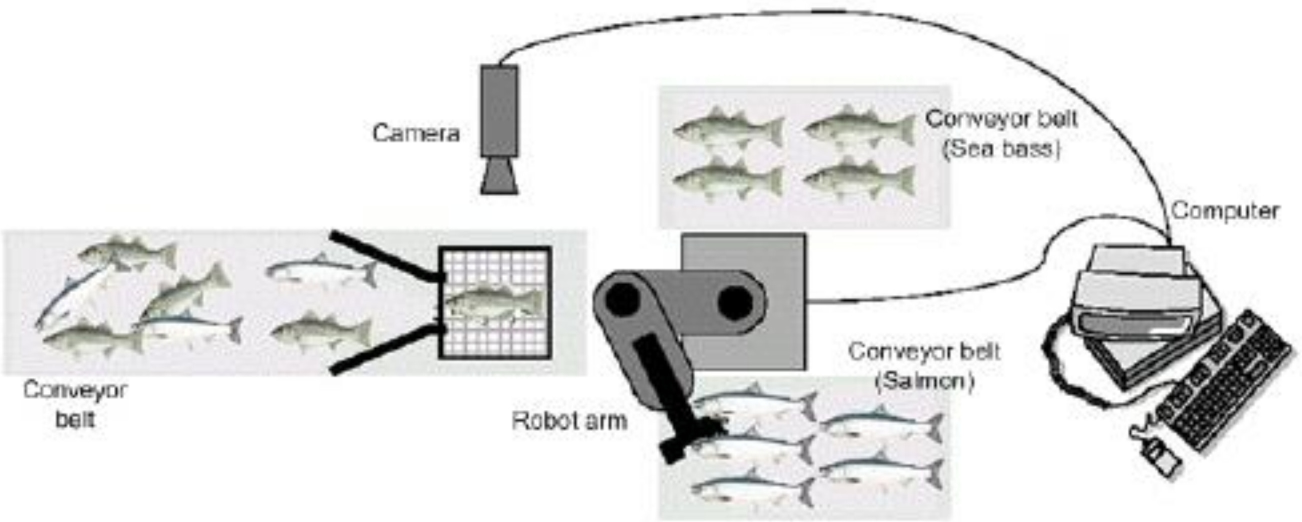
- 针对判断鱼的种类的例子，沿用符号，即X表示为何种鱼的随机变量， L表示为鱼的长度的随机变量。则$P(L|X = 0)$表示鲈鱼的身长取某个值的概率，其光滑过后的概率分布函数可记为$p(L|X = 0)$， $P(L|X = 1)$表示三文鱼的身长取某个值的概率，其光滑过后的概率分布函数可记为$p(L|X = 1)$。

  Taken the fish classification example, let X denote the random variable of fish class, and L for the length of the fish. Then $P(L|X = 0)$ indicates the probability of a specific value of body length for sea bass, and $p(L|X = 0)$ denotes the probability distribution function after smoothing $P(L|X = 0)$. The same deduction applies to the case of salmon.

贝叶斯决策理论
Bayesian Decision Theory

- 现在有了鱼的体长的概率分布函数。则对传送上来的鱼，通过设备取得鱼的体长后，又该如何判断呢？

  Now we have the distribution functions of the fish length of different kinds. Suppose via apparatus we obtained the length of the fish. Then how to decide subsequently?



贝叶斯决策理论
Bayesian Decision Theory

- 我们知道$p(L|X=0)$和$p(L|X=1)$，若同时有$P(X=0)=0.01$，$P(X=1)=0.99$，则如何判断呢？若$P(X=0)=0.3$，$P(X=0)=0.7$呢？

  We already know $p(L|X=0)$ and $p(L|X=1)$. And if further we have $P(X=0)=0.01$, $P(X=1)=0.99$, Then how we make the decision? If $P(X=0)=0.3$ and $P(X=0)=0.7$?

- 我们回忆之前学过的贝叶斯定理（ Bayes Theorem ）：

$$P(B|A)=\frac{P(AB)}{P(A)}=\frac{P(A|B)P(B)}{P(A)}$$

其中$P(B)$为先验概率，$P(A|B)$为条件概率，$P(B|A)$为后验概率，$P(AB)$为联合概率，亦写成$P(A\cap B)$或$P(A,B)$。在一些方法比如朴素贝叶斯的应用中，$P(B)$可能会被抵消掉。

Here $P(B)$ is called prior probability, $P(A|B)$ is called conditional probability, $P(B|A)$ is called posterior probability, $P(AB)$ is called joint probability, sometimes also denoted as $P(A\cap B)$ or $P(A,B)$. In some applications such as naïve Bayesian method, $P(B)$ might be cancelled out.

贝叶斯决策理论
Bayesian Decision Theory

- 若已知鱼体长度的情况下，求鱼为鲈鱼的概率：

  Given the length l, calculate the probability of sea bass:

$$P(X=0|L=l)=\frac{P(X=0\cap L=l)}{P(L=l)}=\frac{P(L=l|X=0)P(X=0)}{P(L=l)}$$

- 若已知鱼体长度的情况下，求鱼为三文鱼的概率：

  Given the length l, calculate the probability of salmon:

$$P(X=1|L=l)=\frac{P(X=1\cap L=l)}{P(L=l)}=\frac{P(L=l|X=1)P(X=1)}{P(L=l)}$$

可以通过比较后验概率的大小判断，定义如下$D(X|L=l)$：

The decision can be made based on the posterior probability. Define $D(X|L=l)$ as follows:

$$D(X|L=l)=\frac{P(X=0|L=l)}{P(X=1|L=l)}=\frac{P(L=l|X=0)P(X=0)}{P(L=l|X=1)P(X=1)}$$

贝叶斯决策理论
Bayesian Decision Theory

- 若$D(X|L=l)>1$，则判断为鲈鱼，否则判断为三文鱼。

  If $D(X|L=l)>1$, Then the output is sea bass, otherwise it is salmon.

- 注意到我们对$P(L|X=0)$或$P(L|X=1)$，均是得到平滑之后的概率分布函数，即$p(L|X=0)$或$p(L|X=1)$，但对$D(X|L=l)$的定义，在有的项是概率分布，有的项是概率分布函数的时候，依然有效，即依然可以依据以下定义的$D(X|L=l)$进行判断：

  It is mentioned here that for $P(L|X=0)$ or $P(L|X=1)$, what we obtained are the distribution functions after smoothing, aka, $p(L|X=0)$ and $p(L|X=1)$. However, in practice, the definition of $D(X|L=l)$ still holds by mixing probability and probability distribution together, just as above. We can still rely on the value of $D(X|L=l)$ to make the decision.

$$D(X|L=l)=\frac{p(L=l|X=0)P(X=0)}{p(L=l|X=1)P(X=1)}$$

神经网络的概率解释
Interpretation of Neural network from the probabilistic perspective

- 考虑用神经网络进行猫和狗的概率分类问题。X表示图片数据，y表示图片辨识结果（标签），均为随机变量；令$f(y|X;w)$表示神经网络，同时看作一个参数化的概率模型（依赖于w），即观察到某个图片输入时，输出为特定类别的概率。令$P(y=0)$表示是猫的图片概率，$P(y=1)$表示是狗的图片的概率；$P(X|y=0;w)$表示当标签是猫时，图片的特征的概率分布； $P(X|y=1;w)$表示当标签是狗时，图片的特征的概率分布；令$P(y=0|X=x;w)$表示当观察到输入为x时，输出是猫的概率，$P(y=1|X=x;w)$表示当观察到输入为x时，输出是狗的概率。

  Consider the application of neural networks for image classification. Let X and y be random variables to denote image data and image prediction (label) respectively. Let $f(y|X;w)$ denote the network, which is treated as a parameterized probabilistic model (by parameter w) simultaneously. It is interpreted as for a given image data observed, the probability of a specific class instance, or a label value. Let $P(y=0)$ and $P(y=1)$ represent the probability of cat and dog images respectively. Now $P(X|y=0;w)$ means the probability distribution of image features given cats, and $P(X|y=1;w)$ means that of dogs. while $P(y=0|X=x;w)$ and $P(y=1|X=x;w)$ indicates the probabilities of cat and dog class labels when observing the input $X=x$.

神经网络的概率解释
Interpretation of Neural network from the probabilistic perspective

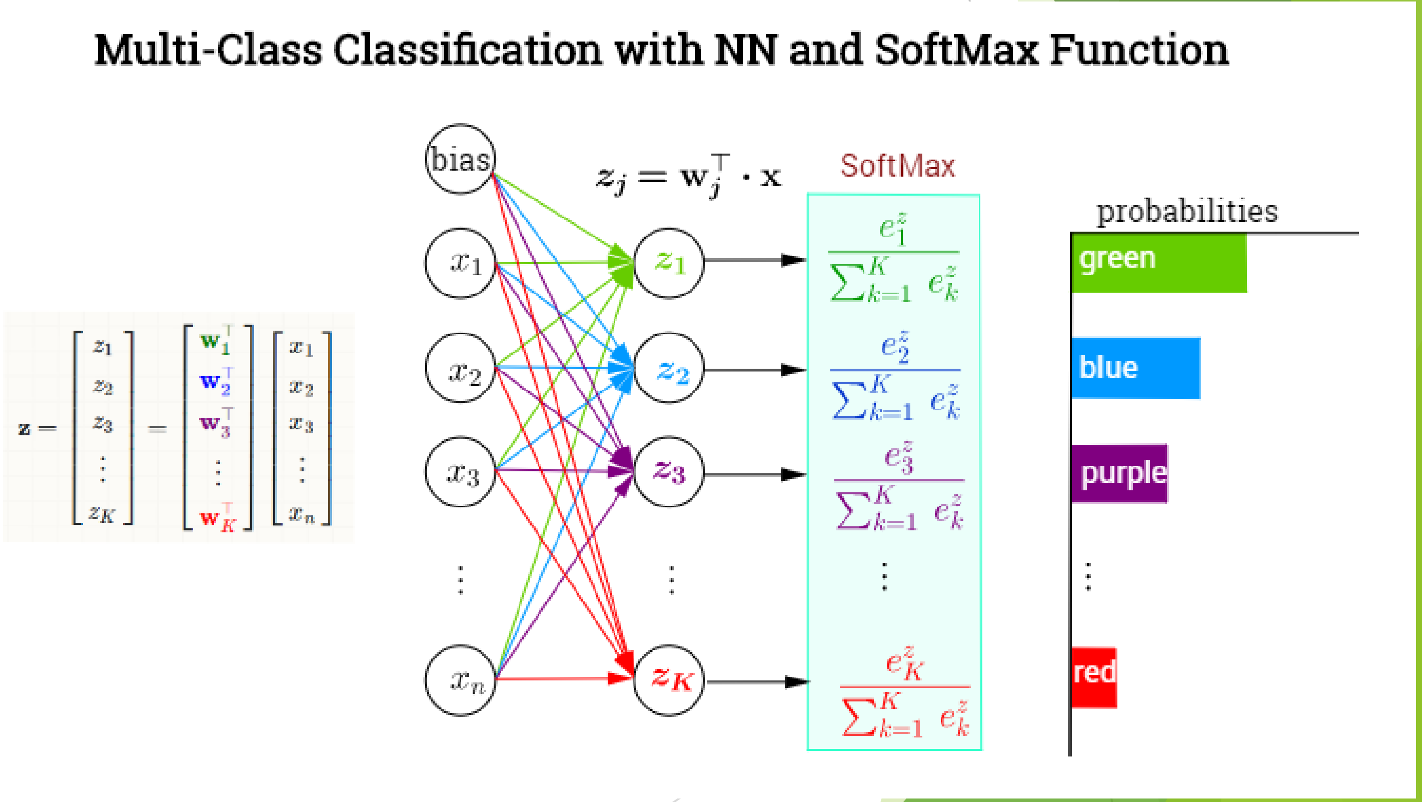
- 为适配上述解释对网络进行的更改：输出概率

  Modification to the neural network to cater to the interpretation: output probability.

- 为适配上述解释对网络进行的定位：估计条件概率

  Retargeting of the neural network to cater to the interpretation: conditional probability estimation.

$$D(y|X=x) = \frac{p(X=x|y=0)P(y=0)}{p(X=x|y=1)P(y=1)}$$



习题

Problems

1. 考虑随机掷两个骰子的情况，如不区分两个骰子，求此时的采样空间。

   Consider the case for tossing two dices, calculate the sampling space (assume there is no distinguishing between two dices).

2. 阅读相关文献，了解什么是独立同分布（i.i.d）随机变量及其在模型简化中的作用。

   Do literature survey to understand what are independently identically distribution random variables and their roles in model simplification.

3. 工程实践中，经常需要从特定分布中采样。但有时构造特定分布并不容易，阅读相关文献，了解如何从简单分布，如均一分布构造特定分布。

   In practice, it might demand to sample from a certain specific distribution. However, to construct a specific distribution is not an easy task. By doing literature survey to find out methods that how to construct complicated distributions from simple distributions such as uniform distribution.

习题
Problems

4. 某个夜晚，一辆出租车肇事后逃逸。该城市共有两家出租车公司，一家公司的出租车均为绿色（“绿色”公司），拥有出租车数量为全市出租车总数的85%；另一家公司的出租车均为蓝色（“蓝色”公司），拥有出租车数量为全市出租车总数的15%。一名目击者称肇事出租车是“蓝色”公司的。法院对目击者的证词进行了测试，发现目击者在出事当时那种情况下正确识别两种颜色的概率是80%。那么肇事出租车是蓝色的概率是多少？

   One night, a taxi fled the scene after an accident. There are two taxi companies in the city, with one company runs all green taxies and another manages all blue ones. The green ones occupy 85% of total account whilst the blue ones constitutes 15%.One witness reported that committed taxi was blue. The court tests the witness’s testimony and find that in that situation, the probability to identify the right color is around 80%. So, in this circumstance what is the chance of the fled taxi is blue?

习题
Problems

- 解答：

根据题目知，

- $P(G) = 0.85$：出租车是绿色的概率；
- $P(B) = 0.15$：出租车是蓝色的概率；
- $P(X = g|G) = 0.80$：当出租车是绿色时，肇事环境下观察到绿色的概率；
- $P(X = g|B) = 0.20$：当出租车是蓝色时，肇事环境下观察到绿色的概率
- $P(X = b|B) = 0.80$：当出租车是蓝色时，肇事环境下观察到蓝色的概率
- $P(X = b|G) = 0.20$：当出租车是绿色时，肇事环境下观察到蓝色的概率

$P(B|X = b) =?$：求当肇事环境下观察到蓝色时，是蓝色出租车的概率

习题
Problems

- Solution：

According to the question，

- $P(G) = 0.85$: The probability that the taxi is green;
- $P(B) = 0.15$: The probability that the taxi is blue;
- $P(X = g|G) = 0.80$: The probability of spotting a green taxi in the accident scene given a green taxi;
- $P(X = g|B) = 0.20$: The probability of spotting a green taxi in the accident scene given a blue taxi;
- $P(X = b|B) = 0.80$: The probability of spotting a blue taxi in the accident scene given a blue taxi;
- $P(X = b|G) = 0.20$: The probability of spotting a blue taxi in the accident scene given a blue taxi;

$P(B|X = b) =?$： To calculate the probability of a blue taxi given the observation of a blue taxi.

习题

Problems

$$P(B|X=b)=\frac{P(B,X=b)}{P(X=b)}=\frac{P(X=b|B)P(B)}{P(X=b)}$$

$$P(X=b)=P(X=b|G)P(G)+P(X=b|B)P(B)$$

$$P(B|X=b)=\frac{P(X=b|B)P(B)}{P(X=b|G)P(G)+P(X=b|B)P(B)}$$

$$=\frac{0.8*0.15}{0.2*0.85+0.8*0.15}=0.41$$