Probabilistic Tools for Machine Learning

Lecture 2

Manuel Balsera

IEOR, Columbia University

January 25, 2018

Outline

Today we will

- Introduce some basic probabilistic notation.
- review statistical inference and maximum likelihood estimation. This should be familiar, at least in the gaussian, linear regression, setting.
- discuss Bayesian inference and how it can be used when data is scarce. We will follow the exposition on [1], but chapters 8 and 9 of [2] covers the same topics and is freely available online.
- introduce the "classical" discrete probability distributions we will use in this course.
- make the connection between statistical inference and supervised machine learning.
- Discuss a couple of loose statistic results that will be useful later
  - hypotesis testing for categorical variables
  - The distribution of the maximum of C random variables.

Conditional Probability

Probability of that random variable X takes value x (the probability density function) will be usually denoted as $P_X(x)$.
We will drop the subscript when the random variable is clear from the context.

Joint Probability $P(Y, X)$:

Probability that both X and Y happen at the same time.
This is a probability distribution over X and Y, normalized as

$$\int \mathrm{d}X \mathrm{d}Y \, P(Y, X) = 1 \tag{1}$$

Conditional Probability $P(Y|X)$

Probability that Y happen given that X has happened.
This is a family of probability distributions over Y (one for each possible value of X) normalized as

$$\int \mathrm{d}Y \, P(Y \mid X) = 1 \tag{2}$$

Marginal Probability P(Y)

Probability of Y ignoring the value of X. This is a single probability distribution

$$\int \mathrm{d}Y \, P(Y) = 1 \tag{3}$$

Bayes Theorem

It is central to inference

Bayes Theorem

$$P(Y, X) = P(Y \mid X)P(X) \tag{4}$$

It relates

- The **join probability** of observing X and Y
- To the **Conditional Probability** of observing Y given X
- and the **Marginal Probability** of observing X.

Independence

Two random variables X and Y are said to be independent if

Independent Random Variables

$$P(Y, X) = P(Y)P(X) \tag{5}$$

or, equivalently, the probability of Y conditional on X is equal to the marginal probability of X

$$P(Y \mid X) = P(Y) \tag{6}$$

We will say that X provides no **information** on the distribution of the values of Y.

Expectation

The expected value of a function f of a random variable X is

Expected Value

$$\bar{f} = \mathbb{E}_X(f) = \int \mathrm{d}x P_X(x) f(x) \tag{7}$$

We will drop the subscript when there is no ambiguity.
Similarly Variance and Covariance are defined as

Variance

$$\mathrm{Var}(f) = \mathbb{E}_X((f - \bar{f})^2) = \int \mathrm{d}x P_X(x) \left(f(x) - \bar{f}\right)^2 \tag{8}$$

Covariance

$$\mathrm{Covar}(f, g) = \mathbb{E}_X(f - \bar{f})(g - \bar{g})) = \int \mathrm{d}x P_X(x) \left(f(x) - \bar{f}\right) \left(g(x) - \bar{g}\right) \tag{9}$$

Central Limit Theorem

Given **independent**, **identically distributed** variables $X_1, X_2, \ldots X_N$ with mean

$$\mathbb{E}(x_i) = \mu \tag{10}$$

and variance

$$\mathrm{Var}(x_i) = \sigma^2 \tag{11}$$

The sum

$$S_N = \sum_{i=1}^{N} X_i \tag{12}$$

has a Gaussian limit distribution

$$S_N \sim \mathcal{N}(N\mu, N\sigma^2) \tag{13}$$

The variables X_i do not need to be Gaussian, just have finite variance.
The sample mean is then

$$\hat{x}_N = \frac{1}{N} S_N = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{14}$$

is also Gaussian with distribution

$$\hat{x}_N = \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}) \tag{15}$$

Maximum Likelihood

Give a set of **independent** observations s_i, $i = 1, \ldots, N$, of random variable $S \in \mathcal{S}$ and a **family** of probability distributions $P(s; \theta)$, parameterized by $\theta \in \Theta$ we define the

Likelihood Function

$$\text{Likelihood}(\theta; \{s_i\}_{i=1}^N) = \prod_{i=1}^N P(s_i; \theta) \tag{16}$$

- This is just the **joint probability** of the data samples seen as a function of θ.
- $\mathcal{S}$ can be arbitrarily complex (text, images, etc). Θ can an also we a complicated set of parameters (SVM, neural network)

It is usual to work with the normalized

Log Likelihood Function

where comes 1/N?

$$\hat{l}(\theta; \{s_i\}) = \frac{1}{N} \sum_{i=1}^N \log P(s_i; \theta) = \frac{1}{N} \sum_{i=1}^N l(\theta; s_i) \tag{17}$$

Maximum Likelihood Estimation

If we wish to estimate the most likely θ given a set of sample data s_i, we use the

Maximum Likelihood Principle

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} P(\{s_i\};\theta) = \arg\max_{\theta} \hat{l}(\theta;\{s_i\}) \tag{18}$$

- $\hat{\theta}_{\mathrm{ML}}$ must satisfy the first order extreme conditions

$$\frac{\partial}{\partial\theta}\hat{l}(\theta,\{s_i\}) = 0 \tag{19}$$

- When $P(S)$ is Gaussian with fixed variance σ and unknown mean θ, this is equivalent to least square estimation of θ.
- When the number of observations N is large, it can be shown that $\hat{\theta}$ will be asymptotically normally distributed around the true θ_0

$$\sqrt{N}(\hat{\theta}_{\mathrm{MLE}} - \theta_0) \sim \mathcal{N}(0, \mathbb{I}^{-1}) \tag{20}$$

where $\mathbb{I}$ is the **Fisher Information Matrix**

$$\mathbb{I}_{j,k} = -\mathbb{E}\left[\frac{\partial^2 l}{\partial\theta_j \partial\theta_k} \mid \theta_0\right] \tag{21}$$

Loss Function

In machine learning is customary to work in terms of the **Loss Function**

$$E(\theta; \{s_i\}) = -\frac{1}{N} \log P(\{s_i\}; \theta) \tag{22}$$

. With this definition the max likelihood estimate is given by

The Maximum Likelihood Optimization Problem

$$\hat{\theta}_{\mathrm{ML}} = \arg\min_{\theta} E(\theta; \{s_i\}) \tag{23}$$

Some times, we will even dispense from probabilistic arguments and just postulate directly a functional form for E.

The Optimization Problem

- Finding the minimum of E is a hard computational problem
- Unfeasible if $\theta \in \Theta \subset \mathbb{R}^D$ is high dimensional ($D \gg 1$) or has a complicated topology (constrains).
- We will be content with finding a **local minimum**.
- For low dimensional θ Newton method, or any of the methods in scipy.optimize will provide good results.
- The complexity of methods that estimate the curvature of the loss function grow at least like D^2, intractable for large D.

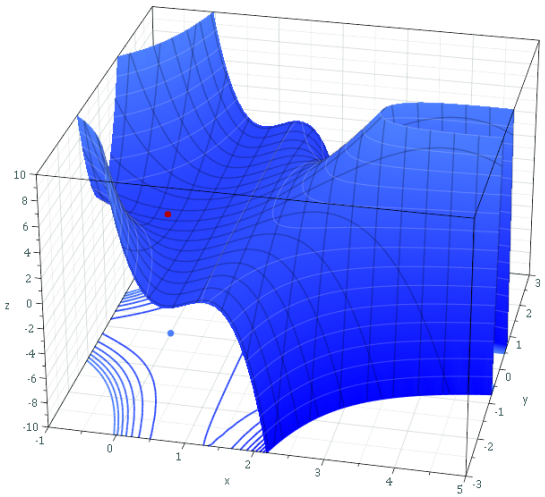


Figure 1: The red dot is a function's local minimum that is not the global minimum. Source: Sam

Gradient Descent

- For large D the workhorse optimization method will be **gradiend descent**.
- We will update parameters in the direction in which the gradient is the steepest.
- We will go **down hill** by a small amount η.
- We will repeat the process a fixed number of times, or until the gradient has vanished

Gradient Descent Algorithm

$$\theta_{t+1} = \theta_t - \eta_t \frac{\partial E}{\partial \theta}(\theta_t) \tag{24}$$

- In general, for each model we will need to know E and $\frac{\partial E}{\partial \theta}$.
- the free paramemeter η is called the **learning rate**.
- If η_t is too large we will **overshoot** and fail to converge.
- If η_t is too small, it will take us a long time to reach the minimum, converge will be **slow**.
- In practice we will **tune** η by trial and error.

Stochastic Gradient Descent

- Gradient descent is a completely **general** algorithm: works for any function E.
- The loss function derived from max likelihood is a sum over N data samples

$$E(\theta; \{s_i\}) = \sum_{i=1}^{N} e(\theta; s_i) \tag{25}$$

where $e(\theta; s) = -\frac{1}{N} \log P(s; \theta)$

Let us choose a small subset B_t of the indexes $i = 1, \ldots N$ (called a **batch**), then

$$E^{B_t} = \sum_{i \in B_t} e_i(\theta; s_i) \tag{26}$$

Stochastic Gradient Descent

$$\theta_{t+1} = \theta_t - \hat{\eta}_t \frac{\partial E^{B_t}}{\partial \theta}(\theta_t) \tag{27}$$

- provided we visit each sample i with the same probability $\frac{\partial E^{B_t}}{\partial \theta}$ is a (stochastic) approximation to $\frac{\partial E}{\partial \theta}$
- Key advantage is that we **don't need all samples together** at the same time.

Bayesian Inference

In Bayesian inference we assume θ is a random variable with **prior** probability distribution $P_0(\theta)$

Bayesian Inference

Posterior Distribution of θ given the observed data $\{s_i\}$

$$P(\theta|\{s_i\}) = \frac{P(\{s_i\}|\theta)P_0(\theta)}{p(\{s_i\})} \tag{28}$$

Predictive Distribution of a new sample point s given the data $\{s_i\}$

$$p(s \mid \{s_i\}) = \int \mathrm{d}\theta \, P(s \mid \theta)P(\theta \mid \{s_i\}) \tag{29}$$

Some terminology

- Likelihood $P(\{s_i\}|\theta)$ as in Max Likelihood Estimation
- Prior $P_0(\theta)$
- Evidence $p(\{s_i\}) = \int \mathrm{d}\theta \, P(\{s_i\}\theta)P_0(\theta)$ this can be hard to compute

Comparison between Max Likelihood and Bayesian Inference

- if we replace the posterior $P(\theta|\{s_i\})$ with its most likely value θ_{MAP} (Maximum A posteriori) we find

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} P(\theta|\{s_i\}) = \arg\max_{\theta} \left\{ \hat{I}(\theta; \{s_i\}) + \frac{1}{N}\Omega(\theta) \right\} \tag{30}$$

  where $\Omega(\theta) = \log P_0(\theta)$
- if we assume an **uninformative prior** $P_0(\theta) = C$ a constant, then $\Omega(\theta) = c$ and the max likehood estimate is the same as the MAP.
- For large sample size N the Bayesian MAP and the Max Likelihood estimate agree with each other
- When N is small the choice of Ω allows for **prior information** to influence our estimates.
- Ω is called a **Regularization** term. It depends on the parameters only, but not on the samples s_i.

Conjugate Priors

- A random variable S is distributed according to a family of probability distributions $P(S;\theta)$
- A family of prior distributions $P_0(\theta;\alpha)$ is said to be conjugate to P if the posterior distribution $P_C(\theta|\{s_i\};\alpha)$ is again of the form $P_0(\theta;\hat{\alpha})$
- The posterior and the prior distributions for θ are in the same parametric family
- Conjugates priors are an analytic **convenience**. There is is no deep reason why prior distributions need to be of conjugate form.
- Wikipedia has a long table of distributions and their conjugate priors.

Distribution	Parameter	Conjugate
Bernouilli	θ	Beta
Gaussian	μ	Gaussian
Poisson	λ	Gamma
Gaussian	σ^2	Inverse Gamma
Multinomial	θ_k	Dirichlet

Table 1: Some common probability distributions an their conjugate priors

Bernoulli Random Variable: Maximum Likelihood Estimates

The simplest random variable (modeling, for example the result of a single coin toss) takes values $s \in 0, 1$. S probability distribution is characterized by a single parameter $0 \leq \theta \leq 1$ and is called

Bernoulli Distribution

$$P(s|\theta) = \theta^s (1-\theta)^{1-s} \tag{31}$$

- It we define $\hat{N}_1 = \sum_i^N s_i$ the log likelihood function is

$$\hat{l}(\theta, \{s_i\}) = \frac{\hat{N}_1}{N} \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta) \tag{32}$$

- and by solving the first order conditions we find the MLE

$$\hat{\theta}_{\text{MLE}} = \frac{\hat{N}_1}{N} \tag{33}$$

- the probability (according to MLE) of the next observation been positive would then be

$$P_{\text{MLE}}(s = 1|\{s_i\}) = \frac{\hat{N}_1}{N} \tag{34}$$

Bernoulli Random Variable: Bayesian Inference with Uninformative Prior

If we do not have any **prior** beliefs about the value of θ, we can assume an uninformative prior

$$P_0(\theta) = 1 \tag{35}$$

for $(0 < \theta < 1)$ with this assumption, if we define $\hat{N}_0 = N - \hat{N}_1$, the number of times that $s = 0$

Posterior Distribution

$$P(\theta|\{s_i\}) = \frac{\theta^{\hat{N}_1}(1-\theta)^{\hat{N}_0}}{B(\hat{N}_1 + 1, \hat{N}_0 + 1)} \tag{36}$$

Predictive Distribution

$$P(s = 1|\{s_i\}) = \frac{\hat{N}_1 + 1}{N + 2}$$

$$P(s = 0|\{s_i\}) = \frac{\hat{N}_0 + 1}{N + 2} \tag{37}$$

The ones added to the empirical counts are known as **pseudo-counts**.
The normalization constant to the posterior distribution is the Beta integral

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} = \int_0^1 \mathrm{d}\theta\, \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{38}$$

Bernoulli's Conjugate Prior

The conjugate Prior to the Bernoulli's distribution is the

Beta Distribution

$$P_0(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{39}$$

- The uninformative prior is $\alpha = 1$, $\beta = 1$.
- the posterior is given by

$$P(\theta \mid \{s_i\}; \alpha, \beta) = \frac{\theta^{\hat{N}_1+\alpha-1}(1-\theta)^{\hat{N}_0+\beta-1}}{B(\hat{N}_1+\alpha, \hat{N}_0+\beta)} \tag{40}$$

- And the predictive distribution is

$$P(s=1 \mid \{s_i\}; \alpha, \beta) = \frac{\hat{N}_1+\alpha}{N+\alpha+\beta} \tag{41}$$

- α and β are **pseudo-counts** and control the strength or our prior beliefs.

Binomial Random Variable

A binomial random variable c_N represents the number of positive outcomes on N independent observations of a Bernoulli random variable.

Binomial Probability Distribution

$$P(c_N = k) = \binom{N}{k} \theta^k (1-\theta)^{N-k} \tag{42}$$

- c_N takes values on the range $\{0, 1, \ldots, N\}$.
- The conjugate prior is a **Beta** distribution as with Bernoulli

Categorical Random Variable

- A categorical variable S can take one of K values with probability

$$P(S = k; \theta) = \theta_k \tag{43}$$

- The categorical distribution is characterized a by K dimensional vector of parameters $(\theta_1, \ldots, \theta_K)$ subject to the constrains

$$\sum_{k=1}^{K} \theta_k = 1; \quad 0 \leq \theta_k \tag{44}$$

- A natural representation of S is as the K-dimensional vector

one-hot-encoding

$$Z(S = k) = (\underbrace{0, 0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{K-k}) \tag{45}$$

- with that representation we can write

$$P(S = k) = \prod_{k=1}^{K} \theta_k^{z_k} \tag{46}$$

Inference for Categorical Random Variables

- The log likelihood function for a categorical variable is given by

$$\hat{l}(\theta; \{s_i\}) = \frac{1}{N} \sum_{i=1}^{N} z_{i,k} \log \theta_k = \sum_k \frac{\hat{N}_k}{N} \log \theta_k \tag{47}$$

  where $\hat{N}_k = \sum_i z_{i,k}$ is the number of observations for which $s_i = k$.
- Optimizing over θ subject to the constrains we find

$$\hat{\theta}_{k,\mathrm{ML}} = \frac{\hat{N}_k}{N} \tag{48}$$

- the conjugate for the categorical distibution is the Dirichlet distribution

$$P_0(\theta) = B(\alpha_1, \alpha_2, \ldots, \alpha_K) \theta_k^{\alpha_k - 1} \tag{49}$$

- and the predictive distribution is then

$$P(s = k \mid \{s_i\}; \alpha) = \frac{\hat{N} + \alpha_k}{N + \sum_k \alpha_k} \tag{50}$$

- the α_k are the **pseudo-counts**, and is common to use a central-Dirichlet distribution where $\alpha_1 = \alpha_2 = \cdots = \alpha_k = \alpha$, so all cases are equally likely.

Univariate Gaussian Random Variable

- A continuous univariate Gaussian variable $S \sim \mathcal{N}(\mu, \sigma^2)$ is parameterized by a two parameter family of distributions

$$P(s; \mu, \sigma^2) = \mathcal{N}(s; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-\mu)^2}{2\sigma^2}} \tag{51}$$

- the sample likelihood function is

$$\hat{l}_g(\mu, \sigma^2; \{s_i\}) = \frac{1}{2N\sigma^2} \sum_{i=1}^{N} (s - \mu)^2 - \frac{1}{2} \log 2\pi\sigma^2. \tag{52}$$

- Optimizing the Maximum Likelihood function is equivalent to solving a **least squares** optimization problem

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{N} \sum_i s_i$$

$$\hat{\sigma^2}_{\mathrm{MLE}} = \frac{1}{N} \sum_i (s_i - \hat{\mu}_{\mathrm{MLE}})^2 \tag{53}$$

Bayesian Inference for Gaussian Variables

1. Known Variance
   - Conjugate prior for mean μ is a Gaussian

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0) \tag{54}$$

   - Posterior distribution for μ is still Gaussian with parameters

$$\hat{\mu}_{PM} = \sigma_N^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i s_i}{\sigma^2} \right) \tag{55}$$

$$\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \tag{56}$$

$$\tag{57}$$

   - Predictive distribution is

$$P(s \mid \{s_i\}; \mu_0, \sigma_0^2) \sim \mathcal{N}(s; \hat{\mu}_{PM}, \sigma^2 + \sigma_N^2) \tag{58}$$

2. Unknown Variance
   - Conjugate Prior for variance is an **inverse Gamma** distribution.
   - predictive distribution for s is a **t-Student** distribution.

Supervised Learning as Statistical Inference

- In a **supervised learning** setting we have some extra structure on the observations

$$s_i = (y_i, x_i) \tag{59}$$

- our goal is to learn $P(y_i|x_i;\theta)$
- the log likelihood function is, as before

$$l(\theta; y, x) = \log P(y \mid x; \theta). \tag{60}$$

- the maximum likelihood estimate is then given by

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \hat{l}(\theta) = \arg\max_{\theta} \left[\frac{1}{N} \sum_{i=1}^{N} l(\theta; y_i, x_i) \right] \tag{61}$$

- Bayesian inference can be done exactly as before by imposing a prior on θ.
- In Machine learning we are particularly concerned with the **predictive** distribution $P(y \mid x, \{y_i, x_i\})$.

The Feature Selection Problem

- Given $y \in \mathcal{Y}$, $x \in \mathcal{X}$ and a family of conditional probability distributions $P(y|x;\theta_0)$ we can apply the statistical inference machinery to learn $\hat{\theta}$.
- We will spend a great deal of the course discussing how to do so.
- a practical problem is: that, given a target (label) space $\mathcal{Y}$ can we select a set of variables $x \in \mathcal{X}$ that are the **most efficient** in determining y?
- Example: y is a person's high in inches. For x we may want to use some or all of
  - age
  - sex
  - zip code
  - income
  - eye color
  - weight
  - first letter of last name.
- This is the **feature selection problem**.
- A natural approach is to look at variables that are **strongly associated** with each other.

A Familiar Example of association

- Let's assume the following relationship between the random variable x and y:

$$y = ax + b + \sigma\epsilon \tag{62}$$

where the noise term is normally distributed: $\epsilon \sim \mathcal{N}(0, 1)$.

- In the language we have been using

$$P(y \mid x) = \mathcal{N}(ax + b, \sigma^2) \tag{63}$$

- Defining $\sigma_Y^2 = \mathrm{Var}(Y)$, $\sigma_X^2 = \mathrm{Var}(X)$ we have the familiar

Correlation Coefficient

$$\rho = \frac{Covar(X, Y)}{\sigma_X \sigma_Y} \tag{64}$$

- the key property of ρ is that the fraction of Ys variance explained by X is proportional to ρ^2

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} \tag{65}$$

Correlation versus Association

- correlation summarized completely the association between X and Y provided that
  1. The relationship between X and Y is linear
  2. The residual distribution of Y given X is normal
- If there is a strong non linear dependency between X and Y correlation coefficient can be a rather poor measure of association.

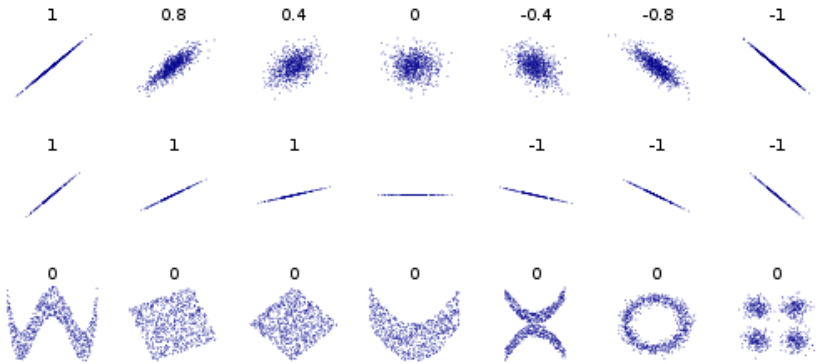


Figure 2: When the relationship between X and Y is non-linear there can be association without correlation. Source:Denis Boigelot, Wikipedia

Distribution test for Categorical Variables

- given N samples s_i, $i = 1, \ldots N$ of a categorical variable $S = 1, \ldots K$ we would like to have a **hypothesis test** to determine if s_i is distributed according to the categorical probabilities θ_k.
- As usual let's define the one-hot encoding of s_i as $z_{i,k}$.
- The expected number of observations of category k is

$$E_k = N\theta_k \tag{66}$$

- The actual observations of category k are

$$\hat{O}_k = \sum_{i=1}^{N} z_{i,k} \tag{67}$$

- the following quantity

$$\sum_k \frac{\left(\hat{O}_k - E_k\right)^2}{E_k} \tag{68}$$

  is distributed as a χ^2 random variable with $K - 1$ degrees of freedom
- We can set a confidence level p and threshold c^2 , such that $P(\chi^2_{K-1} > c^2) = p$.
- If $C^2 > c^2$ we can reject the hypothesis that S is distributed like θ_k.

Association of Categorical Variables

- Given samples (y_i, x_i) of categorical variables $Y = 1, \ldots, K$ and $X = 1, \ldots, D$ we would like some way to quantify how close they are to independence.
- As before let's define the one-hot encoded variables $z_{i,k}^Y$ and $z_{i,d}^X$
- the empirical probabilities of X and Y are given by

$$\hat{p}_k^Y = \frac{1}{N} \sum_i z_{i,k}^Y = \frac{\hat{N}_k^Y}{N} \tag{69}$$

$$\hat{p}_d^X = \frac{1}{N} \sum_i z_{i,d}^X = \frac{\hat{N}_d^X}{N} \tag{70}$$

- if X and Y are independent then the expected and observed number of observations where $y_i = k$ and $x_i = d$ is given by

$$\hat{E}_{k,d} = N \hat{p}_k^Y \hat{p}_d^X \tag{71}$$

$$\hat{O}_{k,d} = \sum_i z_{i,k}^Y z_{i,d}^X = \left(Z^Y\right)^T Z^X \tag{72}$$

Association of Categorical Variables, continued

- The following quantity

$$C^2 = \sum_{k=1}^{K} \sum_{d=1}^{D} \frac{(\hat{O}_{k,d} - \hat{E}_{k,d})^2}{\hat{E}_{k,d}} \tag{73}$$

  is distributed as a χ^2 variable with $(K-1) \times (D-1)$.
- The distribution of C^2 can be use as an **statistical test** of independence.
- more generally, we can use the C^2 value for two categorical variables as a summary of how far they are from independence
- we will use this for **text classification**.

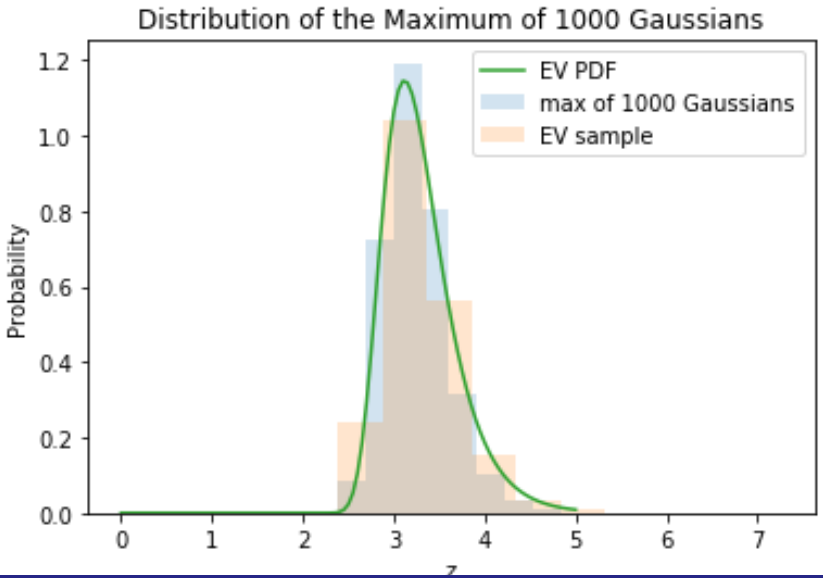
Extreme Value Distribution for Gaussians

Sometimes during a machine learning procedure we will need to pick the **best** one out of C procedures. It is instructive that we study the distribution of the maximum on $C \sim \mathbb{N}(0,1)$ variables.

$$M_C = \max_c(x_1, x_2, \ldots, x_C) \tag{74}$$

It can be shown that

$$\lim_{C\to\infty} P(M_C < z) = e^{e^{-\left(b_C + \frac{z}{b_C}\right)}}, \quad b_C \approx \sqrt{2\log C} \cdots \tag{75}$$

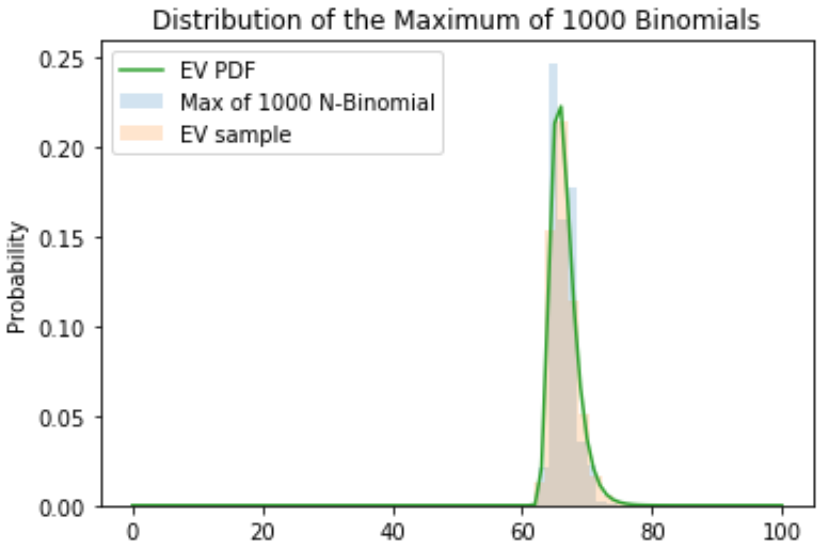


Finding a biased Coin

- Imagine we have C fair coins and a biased one.
- How many coin tosses N do we need to perform to be sure to pickup the biased one with some confidence?
- The outcome of N coin tosses is well approximated by a Gaussian.
- $H_{N,C}$ is number of heads on N tosses of C coins.

$$z = \frac{H_{N,C} - pN}{p(1-p)N} \tag{76}$$

is distributed like M_C, the maximum of C Gaussians.



The Selection Problem

When we select the best among many alternatives we face the

The Selection Problem

The larger the selection set the harder to distinguish true out-performance from random fluctuations.

- This problem is analogous to the asset manager selection problem in finance.
- In ML learning we will have a number of possible methods to chose from, and we will want to determine which one is the best.
- In general, most methods are similar so results are not truly independent, so the asymptotic formulas do not apply.
- In realistic examples we won't able to estimate distributions as with coin tosses, but the phenomenon is the same.

Conclusion

Today we have discussed how to

- make inferences given a family of probability distributions $P(s;\theta)$, and a set of observations $s_i = (y_i, x_i)$, $i = 1, \ldots, N$
- How Bayesian statistics allows us to incorporate our prior beliefs into the estimation process to compensate for limited data.
- In practice the difficulty will to determine the appropriate data space $x_i \in \mathcal{X}$, and how to **choose** a family of functions $P(y \mid x;\theta)$.
- Not all Machine Learning methods are probabilistic. But it is good to keep in mind the estatistical inference example in mind, as we can follow its logic from begining to end.

sectionBibliography

Bibliography


Kevin P. Murphy.
Machine Learning, A Probabilistic Approach.
The MIT Press, 2012.


D. Barber.
Bayesian Reasoning and Machine Learning.
Cambridge University Press, 2012.
http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/091117.pdf.