Machine Learning - Waseda University
Lecture 2: Review of Probability & Statistics

AD

June 2011

Probability Basics

- Begin with a set Ω-the sample space; e.g. 6 possible rolls of a die.
- $\omega \in \Omega$ is a sample point/possible event.
- A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.

  $$0 \leq P(\omega) \leq 1 \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1;$$

  e.g. for a die
  $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6.$
- An event A is any subset of Ω

  $$P(A) = \sum_{\omega \in A} P(\omega)$$

  e.g. $P(\text{die roll} < 3) = P(1) + P(2) = 1/6 + 1/6 = 1/3.$

Random Variables

- A random variable is loosely speaking a function from sample points to some range; e.g. the reals.
- P induces a probability distribution for any r.v. X :

$$P(X = x) = \sum_{\{\omega \in \Omega : X(\omega) = x\}} P(\omega)$$

e.g.
$P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2.$

Why use probability?

- "Probability theory is nothing but common sense reduced to calculation" — Pierre Laplace, 1812.
- In 1931, de Finetti proved that it is irrational to have beliefs that violate these axioms, in the following sense:
  If you bet in accordance with your beliefs, but your beliefs violate the axioms, then you can be guaranteed to lose money to an opponent whose beliefs more accurately reflect the true state of the world.
  (Here, "betting" and "money" are proxies for "decision making" and "utilities".)
- What if you refuse to bet? This is like refusing to allow time to pass: every action (including inaction) is a bet.
- Various other arguments (Cox, Savage).

Probability for continuous variables

- We will also often work with continuous-valued r.v., in this case we use densities .
- We have

$$P(X \in A) = \int_A p(x)\, dx.$$

so

$$\int_\Omega p(x)\, dx = 1$$

- We have

$$P(X \in [x, x + \Delta x]) \approx p_X(x)\, \Delta x$$

- Warning: A density evaluated at a point can be bigger than 1.

Univariate Gaussian (Normal) density

- If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ then the pdf of X is defined as

$$p_X\left(x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

- We often use the precision $\lambda = 1/\sigma^2$ instead of the variance.

Statistics of a distribution

- Recall that the mean and variance of a distribution are defined as

$$\begin{aligned}\mathbb{E}(X) &= \int xp(x)\,dx\\ \mathbb{V}(X) &= \mathbb{E}\left((X-\mathbb{E}(X))^2\right)\\ &= \int (x-\mathbb{E}(X))^2\,p(x)\,dx\end{aligned}$$

- For a Gaussian distribution, we have

$$\mathbb{E}(X)=\mu,\ \mathbb{V}(X)=\sigma^2.$$

- Chebyshev's inequality

$$P(|X-\mathbb{E}(X)|\geq\varepsilon)\leq\frac{\mathbb{V}(X)}{\varepsilon^2}.$$

Law of large numbers

- Assume you have $X_i \overset{\text{i.i.d.}}{\sim} p(\cdot)$ then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) = \mathbb{E}[\varphi(X)] = \int \varphi(x)\, p(x)\, dx$$

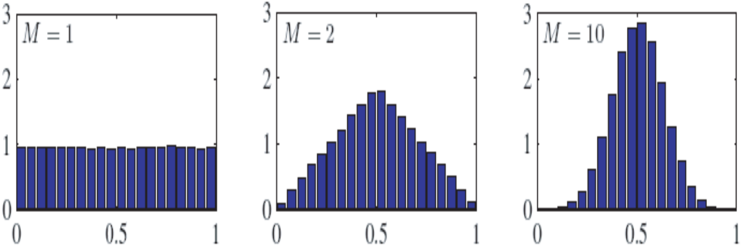
- The result is still valid for weakly dependent random variables; e.g. ergodic Markoc chains.

Central limit theorem

- Let $X_i \overset{\text{i.i.d.}}{\sim} p(\cdot)$ such that $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}(X_i) = \sigma^2$ then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{\text{D}} \mathcal{N}\left(0, \sigma^2\right).$$

- This result is (too) often used to justify the Gaussian assumption



Delta method

- Assume we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)\xrightarrow{\text{D}}\mathcal{N}\left(0,\sigma^2\right)$$

- Then we have

$$\sqrt{n}\left(g\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)-g\left(\mu\right)\right)\xrightarrow{\text{D}}\mathcal{N}\left(0,\left[g'\left(\mu\right)\right]^2\sigma^2\right)$$

- This follows from the Taylor's expansion

$$g\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\approx g\left(\mu\right)+g'\left(\mu\right)\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)+o\left(\frac{1}{n}\right)$$

Prior probability

- Consider the following random variables $\mathit{Sick} \in \{\text{yes,no}\}$ and $\mathit{Weather} \in \{\text{sunny,rain,cloudy,snow}\}$
- We can set a prior probability on these random variables; e.g.

$$\begin{aligned} P\left(\mathit{Sick} = \text{yes}\right) &= 1 - P\left(\mathit{Sick} = \text{no}\right) = 0.1, \\ P\left(\mathit{Weather}\right) &= \left(0.72, 0.1, 0.08, 0.1\right) \end{aligned}$$

- Obviously we cannot answer questions like

$$P\left(\left. \mathit{Sick} = \text{yes} \right| \mathit{Weather} = \text{rainy}\right)$$

as long as we don't define an appropriate distribution.

Joint probability distributions

- We need to assign a probability to each sample point.
- For $(Weather, Sick)$, we have to consider 4×2 events; e.g.

Weather / *Sick*	sunny	rainy	cloudy	snow
yes	0.144	0.02	0.016	0.02
no	0.576	0.08	0.064	0.08

- From this probability distribution, we can compute probabilities of the form $P\left(Sick = \text{yes} \middle| Weather = \text{rainy}\right)$.

Bayes' rule

- We have
$$P(x,y) = P(x|\,y)\,P(y) = P(y|\,x)\,P(x)$$
- It follows that
$$\begin{aligned} P(x|\,y) &= \frac{P(x,y)}{P(y)} \\ &= \frac{P(y|\,x)\,P(x)}{\sum_x P(y|\,x)\,P(x)} \end{aligned}$$
- We have
$$\begin{aligned} P(\mathit{Sick} = \text{yes}|\,\mathit{Weather} = \text{rainy}) &= \frac{P(\mathit{Sick} = \text{yes}, \mathit{Weather} = \text{rainy})}{P(\mathit{Weather} = \text{rainy})} \\ &= \frac{0.02}{0.02 + 0.08} = 0.2 \end{aligned}$$

Example

- Useful for assessing diagnostic prob from causal prob

$$P\left(\text{Cause}\middle|\,\text{Effect}\right) = \frac{P\left(\text{Effect}\middle|\,\text{Cause}\right) P\left(\text{Cause}\right)}{P\left(\text{Effect}\right)}$$

- Let M be meningitis, S be stiff neck:

$$P\left(m\middle|\,s\right) = \frac{P\left(s\middle|\,m\right) P\left(m\right)}{P\left(s\right)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

- Be careful to subtle exchanging of $P\left(A\middle|\,B\right)$ for $P\left(B\middle|\,A\right)$.

Example

- **Prosecutor's Fallacy**. A zealous prosecutor has collected an evidence and has an expert testify that the probability of finding this evidence if the accused were innocent is one-in-a-million. The prosecutor concludes that the probability of the accused being innocent is one-in-a-million.
- This is WRONG.

- Assume no other evidence is available and the population is of 10 million people.
- Defining $A =$ "The accused is guilty" then $P(A) = 10^{-7}$.
- Defining $B =$ "Finding this evidence" then $P(B|A) = 1$ & $P(B|\overline{A}) = 10^{-6}$.
- Bayes formula yields

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} \\ &= \frac{10^{-7}}{10^{-7} + 10^{-6} \times (1 - 10^{-7})} \approx 0.1. \end{aligned}$$

- *Real-life Example*: Sally Clark was condemned in the UK (The RSS pointed out the mistake). Her conviction was eventually quashed (on other grounds).

- You feel sick and your GP thinks you might have contracted a rare disease (0.01% of the population has the disease).
- A test is available but not perfect.
    - If a tested patient has the disease, 100% of the time the test will be positive.
    - If a tested patient does not have the diseases, 95% of the time the test will be negative (5% false positive).
- Your test is positive, should you really care?

- Let A be the event that the patient has the disease and B be the event that the test returns a positive result

$$P\left(A|\,B\right)=\frac{1\times 0.0001}{1\times 0.0001+0.05\times 0.9999}\approx 0.002$$

- Such a test would be a complete waste of money for you or the National Health System.
- A similar question was asked to 60 students and staff at Harvard Medical School: 18% got the right answer, the modal response was 95%!

“General” Bayesian framework

- Assume you have some unknown parameter θ and some data D.
- The unknown parameters are now considered RANDOM of prior density $p(\theta)$
- The likelihood of the observation D is $p(D|\theta)$; i.e. density of the observations
- The posterior is given by

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)}$$

where

$$p(D) = \int p(D|\theta)\,p(\theta)\,d\theta$$

- $p(D)$ is called the marginal likelihood or evidence.

Bayesian interpretation of probability

- In the Bayesian approach, probability describes degrees of belief, instead of limiting frequencies.
- The selection of a prior has an obvious impact on the inference results! However, Bayesian statisticians are honest about it.
- Bayesian inference involves passing from a prior $p(\theta)$ to a posterior $p(\theta|D)$ .We might expect that because the posterior incorporates the information from the data, it will be less variable than the prior.
- We have the following identities

$$\begin{aligned}\mathbb{E}[\theta] &= \mathbb{E}[\mathbb{E}[\theta|D]],\\ \mathbb{V}[\theta] &= \mathbb{E}[\mathbb{V}[\theta|D]] + \mathbb{V}[\mathbb{E}[\theta|D]].\end{aligned}$$

- It means that, *on average (over the realizations of the data X)* we expect the conditional expectation $\mathbb{E}[\theta|D]$ to be equal to $\mathbb{E}[\theta]$ and *the posterior variance to be on average smaller than the prior variance* by an amount that depend on the variations in posterior means over the distribution of possible data.

If (θ, X) are two scalar random variables then we have

$$\mathbb{V}[\theta] = \mathbb{E}[\mathbb{V}[\theta | D]] + \mathbb{V}[\mathbb{E}[\theta | D]].$$

Proof:

$$\begin{aligned}
\mathbb{V}[\theta] &= \mathbb{E}\left(\theta^2\right) - \mathbb{E}(\theta)^2 \\
&= \mathbb{E}\left(\mathbb{E}\left(\theta^2 \middle| D\right)\right) - \left(\mathbb{E}\left(\mathbb{E}(\theta | D)\right)\right)^2 \\
&= \mathbb{E}\left(\mathbb{E}\left(\theta^2 \middle| D\right)\right) - \mathbb{E}\left(\left(\mathbb{E}(\theta | D)\right)^2\right) \\
&\quad + \mathbb{E}\left(\left(\mathbb{E}(\theta | D)\right)^2\right) - \left(\mathbb{E}\left(\mathbb{E}(\theta | D)\right)\right)^2 \\
&= \mathbb{E}\left(\mathbb{V}(\theta | X)\right) + \mathbb{V}\left(\mathbb{E}(\theta | X)\right).
\end{aligned}$$

- Such results appear attractive but one should be careful. Here there is an underlying assumption that the observations are indeed distributed according to $p(D)$.

Frequentist interpretation of probability

- Probabilities are objective properties of the real world, and refer to limiting relative frequencies (e.g. number of times I have observed heads.).
- *Problem*. How do you attribute a probability to the following event "There will be major incidents on the Korean border on the 27th April 2013"?
- Hence in most cases θ is assumed unknown but fixed.
- We then typically compute point estimates of the form $\widehat{\theta} = \varphi(D)$ which are designed to have various desirable quantities when averaged over future data D' (assumed to be drawn from the "true" distribution).

Maximum Likelihood Estimation

- Suppose we have $X_i \overset{\text{i.i.d.}}{\sim} p\left(\cdot\middle|\,\theta\right)$ for $i = 1, ..., N$ then

$$L\left(\theta\right) = p\left(D\middle|\,\theta\right) = \prod_{i=1}^{N} p\left(x_i\middle|\,\theta\right).$$

- The MLE is defined as

$$\widehat{\theta} = \arg\max L\left(\theta\right) = \arg\max l\left(\theta\right)$$

where

$$l\left(\theta\right) = \log L\left(\theta\right) = \sum_{i=1}^{N} \log p\left(x_i\middle|\,\theta\right).$$

MLE for 1D Gaussian

- Consider $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ then

$$p\left(D \middle| \mu, \sigma^2 \right) = \prod_{i=1}^{N} \mathcal{N}\left(x_i; \mu, \sigma^2\right)$$

so

$$l\left(\theta\right) = -\frac{N}{2}\log\left(2\pi\right) - \frac{N}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

- Setting $\frac{\partial l(\theta)}{\partial \mu} = \frac{\partial l(\theta)}{\partial \sigma^2} = 0$ yields

$$\begin{aligned} \widehat{\mu}_{ML} &= \frac{1}{N}\sum_{i=1}^{N} x_i, \\ \widehat{\sigma^2}_{ML} &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \widehat{\mu}_{ML}\right)^2. \end{aligned}$$

- We have $\mathbb{E}\left[\widehat{\mu}_{ML}\right] = \mu$ and $\mathbb{E}\left[\widehat{\sigma^2}_{ML}\right] = \frac{N}{N-1}\sigma^2$.

Consistent estimators

Definition. A sequence of estimators $\widehat{\theta}_N = \widehat{\theta}_N \left(X_1, ..., X_N \right)$ is consistent for the parameter θ if, for every $\varepsilon > 0$ and every $\theta \in \Theta$

$$\lim_{N \to \infty} P_\theta \left(\left| \widehat{\theta}_N - \theta \right| < \varepsilon \right) = 1 \text{ (equivalently } \lim_{N \to \infty} P_\theta \left(\left| \widehat{\theta}_N - \theta \right| \geq \varepsilon \right) = 0).$$

- *Example*: Consider $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, 1\right)$ and $\widehat{\mu}_N = \frac{1}{N} \sum_{i=1}^N X_i$ then $\widehat{\mu}_N \sim \mathcal{N}\left(\mu, N^{-1}\right)$ and

$$P_\theta \left(\left| \widehat{\theta}_N - \theta \right| < \varepsilon \right) = \int_{-\varepsilon\sqrt{N}}^{\varepsilon\sqrt{N}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2} \right) du \to 1.$$

- It is possible to avoid this calculation and use instead Chebychev's inequality

$$P_\theta \left(\left| \widehat{\theta}_N - \theta \right| \geq \varepsilon \right) = P_\theta \left(\left| \widehat{\theta}_n - \theta \right|^2 \geq \varepsilon^2 \right) \leq \frac{\mathbb{E}_\theta \left(\left(\widehat{\theta}_n - \theta \right)^2 \right)}{\varepsilon^2}$$

where $\mathbb{E}_\theta \left(\left(\widehat{\theta}_n - \theta \right)^2 \right) = var_\theta \left(\widehat{\theta}_n \right) + \left(\mathbb{E}_\theta \left(\widehat{\theta}_n - \theta \right) \right)^2.$

- Example of *inconsistent* MLE (Fisher)

$$(X_i, Y_i) \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_i \\ \mu_i \end{array}\right), \left(\begin{array}{cc} \sigma^2 & 0 \\ 0 & \sigma^2 \end{array}\right)\right).$$

- The likelihood function is given by

$$L(\theta) = \frac{1}{(2\pi\sigma^2)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\mu_i)^2 + (y_i-\mu_i)^2\right]\right)$$

- We obtain

$$\begin{aligned} l(\theta) &= cste - n\log\sigma^2 \\ &\quad -\frac{1}{2\sigma^2}\left[2\sum_{i=1}^{n}\left(\frac{x_i+y_i}{2}-\mu_i\right)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-y_i)^2\right]. \end{aligned}$$

- We have

$$\widehat{\mu}_i = \frac{x_i+y_i}{2},\ \widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-y_i)^2}{4n} \rightarrow \frac{\sigma^2}{2}.$$

Consistency of the MLE

- Kullback-Leibler Distance: For any density f, g

$$D(f, g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

- We have

$$D(f, g) \geq 0 \text{ and } D(f, f) = 0.$$

- Indeed

$$-D(f, g) = \int f(x) \log\left(\frac{g(x)}{f(x)}\right) dx \leq \int f(x) \left(\frac{g(x)}{f(x)} - 1\right) dx = 0$$

- $D(f, g)$ is a very useful 'distance' and appears in many different contexts.

Sketch of consistency of the MLE

- Assume the pdfs $f\left(x|\,\theta\right)$ have common support for all θ and $p\left(x|\,\theta\right) \neq p\left(x|\,\theta'\right)$ for $\theta \neq \theta'$; i.e. $S_\theta = \{x : p\left(x|\,\theta\right) > 0\}$ is independent of θ.
- Denote

$$M_N\left(\theta\right) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{p\left(X_i|\,\theta\right)}{p\left(X_i|\,\theta_*\right)}$$

- As the MLE $\widehat{\theta}_n$ maximizes $L\left(\theta\right)$, it also maximizes $M_N\left(\theta\right)$.
- Assume $X_i \overset{\text{i.i.d.}}{\sim} p\left(\cdot|\,\theta_*\right)$. Note that by the law of large numbers $M_N\left(\theta\right)$ converges to

$$\begin{aligned}\mathbb{E}_{\theta_*}\left(\log\frac{p\left(X|\,\theta\right)}{p\left(X|\,\theta_*\right)}\right) &= \int p\left(x|\,\theta_*\right)\log\frac{p\left(x|\,\theta\right)}{p\left(x|\,\theta_*\right)}dx \\ &= -D\left(p\left(\cdot|\,\theta_*\right), p\left(\cdot|\,\theta\right)\right) := M\left(\theta\right).\end{aligned}$$

- Hence, $M_N\left(\theta\right) \approx -D\left(p\left(\cdot|\,\theta_*\right), p\left(\cdot|\,\theta\right)\right)$ which is maximized for θ^* so we expect that its maximizer will converge towards θ_*.

Asymptotic Normality

- Assuming $\widehat{\theta}_N$ is a consistent estimate of θ_*, we have under regularity assumptions

$$\sqrt{N}\left(\widehat{\theta}_N - \theta_*\right) \Rightarrow \mathcal{N}\left(0, \left[I\left(\theta^*\right)\right]^{-1}\right)$$

  where

$$\left[I\left(\theta\right)\right]_{k,l} = -\mathbb{E}\left[\frac{\partial^2 \log p\left(\left. x \right| \theta\right)}{\partial\theta_k \partial\theta_l}\right]$$

- We can estimate $I\left(\theta^*\right)$ through

$$\left[I\left(\theta^*\right)\right]_{k,l} = -\frac{1}{N}\sum_{i=1}^{N} \left.\frac{\partial^2 \log p\left(\left. X_i \right| \theta\right)}{\partial\theta_k \partial\theta_l}\right|_{\widehat{\theta}_N}$$

Sketch of the proof

- We have

$$\begin{aligned} 0 &= l'\left(\widehat{\theta}_N\right) \approx l'(\theta_*) + \left(\widehat{\theta}_N - \theta_*\right) l''(\theta_*) \\ \Rightarrow & \sqrt{N}\left(\widehat{\theta}_N - \theta_*\right) = \frac{\frac{1}{\sqrt{N}} l'(\theta_*)}{-\frac{1}{N} l''(\theta_*)} \end{aligned}$$

- Now $l'(\theta_*) = \sum_{i=1}^{n} \frac{\partial \log p(X_i|\theta)}{\partial \theta}\Big|_{\theta^*}$ where $\mathbb{E}_{\theta_*}\left[\frac{\partial \log p(X_i|\theta)}{\partial \theta}\Big|_{\theta^*}\right] = 0$ and $var_{\theta_*}\left[\frac{\partial \log p(X_i|\theta)}{\partial \theta}\Big|_{\theta^*}\right] = I(\theta_*)$ so the CLT tells us that

$$\frac{1}{\sqrt{N}} l'(\theta_*) \xrightarrow{\text{D}} \mathcal{N}(0, I(\theta_*))$$

where by the law of large numbers

$$-\frac{1}{N} l''(\theta_*) = -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log p(X_i|\theta)}{\partial \theta_k \partial \theta_l}\Big|_{\theta^*} \to I(\theta_*).$$

Frequentist confidence intervals

- The MLE being approximately Gaussian, it is possible to define (approximate) confidence intervals.
- However be careful, "θ has a 95% confidence interval of $[a, b]$" does NOT mean that $P\left(\theta \in [a, b] \middle| D\right) = 0.95$.
- It means

$$P_{D' \sim P(\cdot|D)}\left(\theta \in \left[a\left(D'\right), b\left(D'\right)\right]\right) = 0.95$$

i.e., if we were to repeat the experiment, then 95% of the time, the true parameter would be in that interval. This does not mean that we believe it is in that interval given the actual data D we have observed.

Bayesian statistics as an unifying approach

- In my opinion, the Bayesian approach is much simpler and more natural than classical statistics.
- Once you have defined a Bayesian model, everything follows naturally.
- Defining a Bayesian model requires specifying a prior distribution, but this assumption is clearly stated.
- For many years, Bayesian statistics could not be implemented for all but the simple models but now numerous algorithms have been proposed to perform approximate inference.