PACS Part 2, Lecture 5

The Normal Distribution

- the normal distribution's density is $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
- setting $I = \int_{-\infty}^{\infty} e^{-x^2/2} dx$, we have $I^2 = \int_{\mathbb{R}^2} e^{-\|x\|^2/2} dx$ by Fubini's theorem
- using the substitution $\psi(r, \alpha) = (r \cos \alpha, r \sin \alpha)$, we get

$$I^2 = \int_0^\infty \int_0^{2\pi} e^{-r^2/2} |\det D\psi(r, \alpha)| d\alpha dr = \int_0^\infty \int_0^{2\pi} e^{-r^2/2} r d\alpha dr = 2\pi$$

- thus $I = \sqrt{2\pi}$, and ϕ is indeed the density of a probability measure
- by symmetry of ϕ around 0, we have $\mathbb{E}X = 0$
- for the variance, we have

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx - \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} \Big|_{-\infty}^{\infty} = 1$$

  using integration by parts
- this justifies the notation $\mathcal{N}(0, 1)$ for the standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = \sqrt{\mathrm{Var}(X)} = 1$
- the general form of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and standard deviation σ has density $\frac{1}{\sqrt{2\pi}\sigma} e^{-((x-\mu)/\sigma)^2/2}$
- if $X \sim \mathcal{N}(\mu, \sigma^2)$, then setting $Z = (X - \mu)/\sigma$ gives

$$\begin{aligned} \mathbb{P}(Z \leq z) = \mathbb{P}(X \leq \sigma z + \mu) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\sigma z + \mu} e^{-((t-\mu)/\sigma)^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx \end{aligned}$$

- hence $Z \sim \mathcal{N}(0, 1)$, which means that $\mathbb{E}X = \sigma \mathbb{E}Z + \mu = \mu$ and $\mathrm{Var}(X) = \mathrm{Var}(\sigma Z) = \sigma^2 \mathrm{Var}(Z) = \sigma^2$

Chernoff Bound for the Normal Distribution

- the normal distribution is tightly concentrated around its mean, which can be shown by a Chernoff bound
- the MGF of $X \sim \mathcal{N}(\mu, \sigma^2)$ is:

$$M_X(t) = \mathbb{E}e^{tX} = e^{t^2\sigma^2/2 + \mu t}$$

- using the MGF, we can show that the sum of two independent normally distributed random variables is itself normally distributed: $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ if $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent
- applying the Chernoff bound twice with the optimal parameter $t = a$ and the union bound once, we get

$$\mathbb{P}\left(\left|\frac{X-\mu}{\sigma}\right| \geq a\right) \leq 2e^{-a^2/2}$$

  if $X \sim \mathcal{N}(\mu, \sigma^2)$

Central Limit Theorem

- the central limit theorem states that the averages of i.i.d. random variables are approximately normally distributed as the number of samples grows
- it does *not* claim that random variables themselves are normally distributed
- let $X_1, X_2, \ldots$ be i.i.d. with finite expected value μ and finite variance σ^2, and set $\bar{X}_n = \sum_{i=1}^n X_n/n$
- the law of large numbers states that $\bar{X}_n \to \mu$ almost surely
- the central limit theorem is a more precise version:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \in [a, b)\right) = \Phi(b) - \Phi(a)$$

  for all $a \leq b$ where Φ is the cumulative distribution function of $\mathcal{N}(0, 1)$
- Let us check convergence of the MGFs: Set $Z_i = (X_i - \mu)/\sigma$ and $Y_n = \sum_{i=1}^n Z_i/\sqrt{n}$. We have $\mathbb{E}e^{tZ_i/\sqrt{n}} = M_Z(t/\sqrt{n})$ where $M_Z(t) = M_{Z_i}(t)$. It follows that:

$$M_{Y_n}(t) = \left(M_Z(t/\sqrt{n})\right)^n$$

  We want to show that $M_{Y_n}(t) \to e^{t^2/2}$. For that, we set $L(t) = \log M_Z(t)$ and show, using L'Hôpital's rule, that:

$$\begin{aligned}\lim_{n\to\infty} nL(t/\sqrt{n}) &= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})n^{-3/2}t/2}{-n^{-2}} = \lim_{n\to\infty} \frac{L'(t/\sqrt{n})t}{2n^{-1/2}} \\ &= \lim_{n\to\infty} \frac{-L''(t/\sqrt{n})n^{-3/2}t^2/2}{-n^{-3/2}} = \lim_{n\to\infty} L''(t/\sqrt{n})\frac{t^2}{2} = \frac{t^2}{2}\end{aligned}$$

  where we used $L(0) = 0$, $L'(0) = \mathbb{E}Z_i = 0$, and $L''(0) = \mathbb{E}Z_i^2 = 1$.

Example: Opinion Polls

- assume yes/no opinions that are i.i.d. Bernoulli variables with parameter p

- we are looking for a confidence interval $[\tilde{p}-\delta, \tilde{p}+\delta]$ with $\mathbb{P}(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) \geq 1-\gamma$
- let's choose $\gamma = \delta = 0.05$
- how many samples do we need?
- often-made but potentially dangerous assumption: $\bar{X}_n$ is normally distributed with mean $\mathbb{E}X_i = p$ and variance $\mathrm{Var}(X_i)/n = p(1-p)/n$
- we are then looking for n such that

$$\mathbb{P}\left(\left|\bar{X}_n - p\right| \geq \delta\right) = 2\left(1 - \Phi\left(\frac{\sqrt{n}\delta}{\sqrt{p(1-p)}}\right)\right) \leq \gamma = 0.05$$

- looking up the corresponding argument of Φ, we have to solve

$$\frac{\delta\sqrt{n}}{\sqrt{p(1-p)}} \geq 1.96,$$

  that is,

$$n \geq 385 \geq \left(20 \cdot 1.96 \cdot \sqrt{p(1-p)}\right)^2$$

  suffices, where we used $p(1-p) \leq 1/4$

Maximum-Likelihood Estimators

- given a parameterized family of probability distributions and a set of i.i.d. samples, we seek to estimate the parameters
- in case of a discrete random variable X, the maximum-likelihood estimator (MLE) is the parameter θ that maximizes

$$\prod_{i=1}^{n} \mathbb{P}_\theta(X = x_i)$$

- in case of a continuous random variable X, it maximizes

$$\prod_{i=1}^{n} f_\theta(x_i)$$

  where f_θ is the probability density with parameter θ
- example: for a Bernoulli random variable, the MLE of the parameter p is $p = k/n$ where k is the number of successes among the n samples
- example: for a normal distribution, the MLE of the parameters μ and σ are $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

- every estimator can be itself seen as a random variable when fixing the ground-truth probability distribution
- an estimator Θ_n that considers n samples is *unbiased* if $\mathbb{E}\Theta_n = \theta$ where θ is the true value of the estimated parameter
- it is *asymptotically unbiased* if $\mathbb{E}\Theta_n \to \theta$ as $n \to \infty$
- the sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is always an unbiased estimated of the expected value of the X_i
- the sample variance $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2$ is only asymptotically unbiased:
$$\mathbb{E}S_n^2 = \frac{n-1}{n}\mathrm{Var}(X_i)$$

Expectation–Maximization Algorithm

- we seek to estimate the parameters $\theta = (\gamma, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ of a mixture of two normal distributions
- a sample is chosen according to $\mathcal{N}(\mu_1, \sigma_1^2)$ with probability γ, and according to $\mathcal{N}(\mu_2, \sigma_2^2)$ with probability $1 - \gamma$.
- analytically calculating the MLE in infeasible
- the Expectation–Maximization (EM) algorithm for this problem iterates the Expectation step followed by the Maximization step:
  1. for every sample x_i, compute the conditional probabilities $p_1(x_i)$ and $p_2(x_i)$ that it was sampled according to the first or the second normal distribution given, using the current parameter values
  2. update the parameters μ_j and σ_j that maximize the expected likelihood according to the $p_j(x_i)$, and γ to the average of the $p_1(x_i)$
- this algorithm is not guaranteed to convergence to the global maximum, but it will approach a local maximum:
$$\begin{aligned} L(x, \gamma^t, \mu_1^t, \mu_2^t, \sigma_1^t, \sigma_2^t) &\le L(x, \gamma^{t+1}, \mu_1^t, \mu_2^t, \sigma_1^t, \sigma_2^t) \\ &\le L(x, \gamma^{t+1}, \mu_1^{t+1}, \mu_2^{t+1}, \sigma_1^{t+1}, \sigma_2^{t+1}) \end{aligned}$$