Basic Probability Cheat Sheet

September 20, 2018

1 Probability and Expectation

1.1 Bayes Rule

Bayes rule:

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{\int p(\mathbf{X}|\theta)p(\theta)\mathrm{d}\theta} = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

If $\mathbf{X}$ represents data and θ is an unknown quantity of interest, the Bayes rule can be interpreted as making *inference* about θ based on the data $\mathbf{X}$ (Bayesian inference) in the form of the posterior distribution $p(\theta|\mathbf{X})$.

Remark. In the machine learning course, you will encounter the words 'learning' and 'inference'. From a Bayesian point of view, there's no difference between those two (because everything is expressed by posteriors) . But machine learning people tend to use 'learning' as tuning parameters of a model using data and 'inference' as computing some quantity with the model (sometimes this includes evaluating a posterior distribution). This distinction is not exhaustive but may be good to know to avoid confusion.

1.2 Some Useful Formulas of Conditional Expectations

- $\mathbb{E}[X] = \mathbb{E}_Y \left[\mathbb{E}_{X|Y}[X|Y]\right]$
- $\mathrm{Var}[X] = \mathrm{Var_Y} \left[\mathrm{\mathbb{E}_{X|Y}[X|Y]}\right] + \mathbb{E}_Y \left[\mathrm{Var_{X|Y}[X|Y]}\right]$

2 Asymptotic Theory

Theorem. ***The Law of Large Numbers*** *Let X_1,$X_2 \ldots$, be independent identically distributed (i.i.d.) real random variables. Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}X_1$. If $\mathbb{E}\,|X_1| < \infty$, then $S_n \to \mu$ as $n \to \infty$*[1].

[1] To be precise, we need to define convergence of random variables.

Theorem. ***The Central Limit Theorem*** *Let $X_1, X_2 \dots,$ be as above. Let $\sigma^2 = \mathrm{Var}[X_1]$. Under a (stronger) assumption $\mathbb{E}X_1^2 < \infty$, the probability distribution of $\sqrt{n}\frac{(S_n - \mu)}{\sigma}$ converges to the standard normal distribution $\mathcal{N}(0,1)^2$.*

3 Miscellaneous

- Linearity. Let X obeys a multivariate normal distribution. $\mathcal{N}(\mu, \Sigma)$. Then, $AX \sim \mathcal{N}\left(A\mu, A\Sigma A^\top\right)$, where A is a matrix of appropriate shape.

- Product of normal densities. Let $N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(x-\mu)^2/(2\sigma^2)\right)$, then

$$\mathcal{N}(x;\mu_1,\sigma_1^2)\mathcal{N}(x;\mu_2,\sigma_2^2) = \mathcal{N}(\mu_1;\mu_{2,}\sigma_1^2+\sigma_2^2)\mathcal{N}\left(x;\frac{\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}},\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)$$

[2] Assume $\sigma = 1$ for simplicity. In contrast with the law of large numbers, what the central limit theorem says is that if you multiply the error of the estimate of the mean $S_n - \mu$ by $\sqrt{n}$, the distribution of the amplified error $\sqrt{n}(S_n - \mu)$ is a Gaussian $\mathcal{N}(0,1)$ for sufficiently large n. If you don't, the error converges to a point (zero) as the variance tends to 0 , which agrees with the law of large numbers.