Machine Learning

Homework 1

Ari Feiglin

1.1 Exercise

(1) Let X, Y be independent, show that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

(2) Prove that $\operatorname{Cov}(X, Y) = 0$.

(3) Give an example of uncorrelated RVs which are not independent.

(1) By LOTUS (the law of the unconscious statistician), since $p_{X,Y} = p_X \cdot p_Y$:

$$\mathbb{E}[XY] = \int_{\mathbb{R}^2} xyp_{X,Y}(x,y) = \int_{\mathbb{R}^2} xyp_X(x)p_Y(y) = \int_{\mathbb{R}} xp_X(x)\left(\int_{\mathbb{R}} yp_Y(y)\right) = \int_{\mathbb{R}} xp_X(x) \cdot \int_{\mathbb{R}} yp_Y(y) = \mathbb{E}[X]\,\mathbb{E}[Y]$$

(2) By definition

$$\operatorname{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

since $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$, this is zero.

(3) Let X be symmetric (i.e. its pdf p_X is even), $Y = f(X)$ so that Y is not independent of X. Then

$$\mathbb{E}[X] = \int_{\mathbb{R}} xp_X(x) = 0$$

since xp_X is odd. Thus

$$\operatorname{Cov}(X, Y) = \mathbb{E}[XY] = \int_{\mathbb{R}^2} xyp_{x,y}(x,y) = \int_{\mathbb{R}} xf(x)p_X(x)$$

since $p_{x,y}(x,y) = 0$ if $y \neq f(x)$. So if $f(x)$ is even, then $xf(x)p_X(x)$ is odd, so this integral becomes zero. So let us take $X \sim U(-1, 1)$ and $Y = X^3$.

1.2 Exercise

What's the best choice in the Monty Hall problem?

Let $X \in \{1, 2, 3\}$ be the door chosen, $Y \in \{1, 2, 3\}$ be the door with the prize. We know that $Z \sim U\{1, 2, 3\}$. Now we ask what the probability of $Z = X$ is modulo that a door was opened without the prize. If $Z = X$, then the door opened could be any of the two and this gives no extra information, so $\mathbb{P}(Z = X) = \frac{1}{3}$. If $Z \neq X$, then the door opened is already determined, but now $\mathbb{P}(Z \neq X) = \frac{2}{3}$. So there is a higher probability that $Z \neq X$, so the best strategy is to change choices.

1.3 Exercise

Let $x_1, \dots, x_n \in \mathbb{R}^d$, and let $C = \sum_{i=1}^n x_i x_i^\top$.

(1) Show that C is positive semi-definite.

(2) Prove that C has rank at most $\min\{n, d\}$.

(1) Let $u \in \mathbb{R}^d$ then

$$u^\top Cu = \sum_{i=1}^n u^\top x_i x_i^\top u = \sum_{i=1}^n (x_i u)^\top (x_i u) = \sum_{i=1}^n \|x_i u\|^2 \geq 0$$

so C is positive semi-definite.

(2) We know that $\operatorname{rank}(A+B) \leq \operatorname{rank} A + \operatorname{rank} B$ so $\operatorname{rank} C \leq \sum_{i=1}^n \operatorname{rank}\big(x_i x_i^\top\big)$. Furthermore $\operatorname{rank}(AB) \leq \operatorname{rank} A, \operatorname{rank} B$ so $\operatorname{rank}\big(x_i x_i^\top\big) \leq 1$, thus $\operatorname{rank} C \leq n$. Furthermore $C \in \mathbb{R}^{d\times d}$ so $\operatorname{rank} C \leq d$ as required.

1.4 Exercise

Show that if a jointly normal distributions are uncorrelated, then they are independent.

Suppose $X \sim \mathcal{N}(\mu, \Sigma)$ so the pdf of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}^n} \cdot \frac{1}{\sqrt{\det \Sigma}} \exp\left(\frac{-(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right)$$

Let $\Sigma = \operatorname{diag}[\sigma_1, \ldots, \sigma_n]$ then

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = \sum_{i=1}^n \sigma_i^{-1}(x_i - \mu_i)^2$$

So since $\det \Sigma = \sigma_1 \cdots \sigma_n$,

$$f_X(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\sigma_i}} \cdot \exp\left(\frac{-(x_i-\mu_i)^2}{2\sigma_i}\right)$$

Since $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$ (since $\operatorname{Var}(X_i) = \Sigma_{ii} = \sigma_i$), this is just

$$f_X(x) = \prod_{i=1}^n f_{X_i}(x_i)$$

as required.

1.5 Exercise

Let $X \sim \mathcal{N}(0, I)$ of dimension d.

(1) For $d = 1$, what is the probability of a sample landing within one standard deviation of the mean?

(2) For $d \to \infty$ show that this probability tends to zero.

(1) This is just $\mathbb{P}(-1 \leq Z \leq 1) = \mathbb{P}(Z \leq 1) - \mathbb{P}(Z \leq -1) = 0.84134 - 0.15866$.

(2) $\{\|X\| \leq 1\}$ is contained in the set that $|X_i| \leq 1$ for every $i \in [d]$. The probability of each of these is $c \in (0,1)$ for the c computed in the previous subquestion. Since X_i are all independent (by the previous question/definition), we have that

$$\mathbb{P}\left(\bigwedge_i |X_i| \leq 1\right) = \prod_{i=1}^d \mathbb{P}(|X_i| \leq 1) = c^d \longrightarrow 0$$

since $c^d \longrightarrow 0$ for $c \in (-1, 1)$.

1.6 Exercise

Let X be a d-dimensional random vector with covariance matrix Σ.

(1) Let $w \in \mathbb{R}^d$, and let $y = w^\top X$. Show that $\operatorname{Var}(y) = w^\top \Sigma w$.

(2) Prove that Σ is positive semi-definite.

(1) We know that

$$\operatorname{Var}(y) = \operatorname{Var}\left(\sum_i w_i X_i\right) = \sum_i w_i^2 \operatorname{Var}(X_i) + 2\sum_{i<j} w_i w_j \operatorname{Cov}(X_i, X_j) = \sum_i w_i^2 \operatorname{Var}(X_i) + \sum_{i \neq j} w_i 2_j \operatorname{Cov}(X_i, X_j)$$

and

$$w^\top \Sigma w = \sum_{i,j} w_i \Sigma_{ij} w_j = \sum_{i,j} w_i w_j \operatorname{Cov}(X_i, X_j) = \sum_i w_i^2 \operatorname{Var}(X_i) + \sum_{i \neq j} w_i w_j \operatorname{Cov}(X_i, X_j) = \operatorname{Var}(y)$$

as required.

(2) Let $w \in \mathbb{R}^d$ and define $y = w^\top X$, then $w^\top \Sigma w = \operatorname{Var}(y) \geq 0$, so Σ is positive semi-definite.