Summer of ML 2021: Week 1 Day 1 Quiz
Linear Algebra and Probability

by Nirjhar Das

Tuesday 15$^{\text{th}}$ June, 2021

Questions

1. A is a square matrix. $tr(A) = 2$. $A^2 = A$. What is the rank of A?

   A. 1

   B. 2

   C. Depends on size of A

2. X is a random variable sampled from a Normal Distribution i.e. $X \sim N(0, 1)$. Then $\Pr(X = 0) =$?

   A. $\frac{1}{\sqrt{2\pi}}$

   B. 1

   C. 0

3. X_1 and X_2 are two independent random variables drawn from Poisson distribution [denoted as $P(\lambda)$ where λ is the mean]. Let $X_1 \sim P(1)$ and $X_2 \sim P(2)$. Let $Z = X_1 + X_2$. Then $Z \sim$?

   A. $N(3, \sqrt{5})$, where $N(\mu, \sigma)$ denote Normal Distribution with mean μ and variance σ^2

   B. $P(3)$

   C. $P(\sqrt{5})$

4. Which of the following are TRUE regarding PCA?

   A. In the reduced dimension, all the information present in the actual data (higher dimension) is present intact.

   B. The features (principal components) obtained by PCA are independent of each other.

   C. PCA is translation invariant (shifting data doesn't change in principal components) but not scale invariant (scale affects PCs)

   D. The low dimensional features are interpretable, that is, their meaning can be explained.

5. Which are TRUE regarding Naive Bayes classifier?

   A. It assumes conditional independence of data features.

   B. It assumes low importance of features that have low conditional probabilities.

   C. It assumes total independence of data features.

   D. It gives equal importance to all features.

Answers

1. **Answer: B**
Since A is a square matrix, we can write $A = Q \cdot \Sigma \cdot Q^T$ by Eigendecomposition, where Q is orthonormal and hence $Q \cdot Q^T = Q^T \cdot Q = I$ and Σ is a diagonal matrix with diagonal entries as eigenvalues λ_i's of A.
Thus $A^2 = (Q \cdot \Sigma \cdot Q^T) \cdot (Q \cdot \Sigma \cdot Q^T) = Q \cdot \Sigma \cdot Q^T \cdot Q \cdot \Sigma \cdot Q^T = Q \cdot \Sigma \cdot I \cdot \Sigma \cdot Q^T = Q \cdot \Sigma^2 \cdot Q^T$. Now note that

$$\Sigma = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots \\ 0 & \lambda_2 & 0 & \dots \\ \vdots & \vdots & \ddots & \end{bmatrix} \implies \Sigma^2 = \begin{bmatrix} \lambda_1^2 & 0 & 0 & \dots \\ 0 & \lambda_2^2 & 0 & \dots \\ \vdots & \vdots & \ddots & \end{bmatrix}$$

But we have $A^2 = A \implies \Sigma^2 = \Sigma \implies \lambda_i^2 = \lambda_i \quad \forall\, i = 1, 2 \dots n$. Thus $\lambda_i = 0$ *or* $1 \; \forall \; i = 1, 2 \dots n$.
Also, we have $tr(A) = 2$ but $tr(A) = \sum_{i=1}^{n} \lambda_i \implies \sum_{i=1}^{n} \lambda_i = 2$. Thus, A has only two non-zero eigenvalues $\because \lambda_i = 0$ *or* $1 \; \forall \; i$. Hence $rank(A) = 2$.

2. **Answer: C**
Note that X is a **continuous** random variable. Hence, *probability density function* (pdf) $f_X(x)$ is defined as:

$$\Pr(a \le X \le b) = \int_a^b f_X(x)\, dx$$

Thus $\Pr(X = a)$ can be written as $\lim_{\epsilon \to 0} \Pr(a \le X \le a + \epsilon)$ which is equal to

$$\lim_{\epsilon \to 0} \int_a^{a+\epsilon} f_X(x)\, dx = 0 \; for \; all \; finite \; f_X(x)$$

In short, $\Pr(X = a) = 0$ **for all continuous random variable** X**.**

3. **Answer: B**
We know for Poisson distribution $P(\lambda)$,

$$\Pr(X = k) = \frac{\lambda^k \, e^{-\lambda}}{k!}$$

Now, $\Pr(Z = k) = \Pr(X_1 + X_2 = k) = \sum_{i=0}^{k} \Pr(X_1 = i,\ X_2 = k - i)$
But X_1 and X_2 are independent variables and hence, $\Pr(X_1 = a,\ X_2 = b) = \Pr(X_1 = a) \cdot \Pr(X_2 = b)$
Thus

$$\begin{aligned} \Pr(Z = k) &= \sum_{i=0}^{k} \Pr(X_1 = i) \cdot \Pr(X_2 = k - i) \\ &= \sum_{i=0}^{k} \frac{\lambda_1^i \, e^{-\lambda_1}}{i!} \cdot \frac{\lambda_2^{(k-i)} \, e^{-\lambda_2}}{(k-i)!} \\ &= \frac{1}{k!} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \lambda_1^i \, \lambda_2^{(k-i)} \, e^{-(\lambda_1 + \lambda_2)} \\ &= \frac{e^{-(\lambda_1 + \lambda_2)}}{k!} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \lambda_1^i \, \lambda_2^{(k-i)} \\ &= \frac{e^{-(\lambda_1 + \lambda_2)}}{k!} \, (\lambda_1 + \lambda_2)^k \end{aligned}$$

Thus, $Z \sim P(\lambda_1 + \lambda_2)$

4. **Answer: B, C**

A. This false because we discard several non-zero eigenvalues and their corresponding eigenvectors to obtain the low dimensional representation. Thus information is certainly lost. In actual practice, only about 90% of the information (variance) is retained while 10% is discarded.

B. The principal components (eigenvectors of the covariance matrix) are orthogonal to each other and hence linearly independent.

C. $Cov(x,\ y) = \sum_{i=1}^{m}(x_i - \bar{x}) \cdot (y_i - \bar{y})$. Thus, the shift $x_i \to x_i + c$ doesn't affect $x_i - \bar{x}$ where c is a constant. Thus PCA is shift invariant. However, the transformation $x_i \to \frac{x_i}{c}$ affects the covariance as $Cov(x',\ y) = \sum_{i=1}^{m}(x_i' - \bar{x'}) \cdot (y_i - \bar{y}) = \sum_{i=1}^{m}(\frac{x_i - \bar{x}}{c}) \cdot (y_i - \bar{y}) = \frac{1}{c} \cdot Cov(x,\ y) \neq Cov(x,\ y)$. Thus scaling affects PCA.

D. The low dimensional features obtained using PCA cannot be given direct physical interpretation as they are composed of principal components which are aligned along the directions of greatest variances, which in turn has no physical meaning.

5. **Answer: A, D**

A. The Naive Bayes classifer assumes that $\Pr(x_i|c)$ are independent $\forall\ i = 1, 2, \ldots n$ where x_i's are features and c is the class. However, it makes no assumption that $\Pr(x_i)$'s are independent. Thus only conditional probabilities are assumed to be independent and not the total probabilities.

B. Refer to part D.

C. Refer to part A.

D. Naive Bayes predicts class as:

$$\Pr(C = c \,|\, X_1 = x_1,\ X_2 = x_2, \ldots X_n = x_n) = \frac{\Pr(C = c)(\prod_{i=1}^{n} \Pr(X_i = x_i \,|\, C = c))}{\Pr(X_1 = x_1,\ X_2 = x_2, \ldots X_n = x_n)}$$

Observe that the denominator of the *R.H.S* is equal for all classes c. Now, in the numerator, the conditionally probability of every feature is multiplied together without any preference for or against a particular feature. Thus, Naive Bayes classifier gives equal importance to all features. This is in contrast to Linear Regression, which determines the weight (importance) of each feature in predicting the target.