Probabilistic ML IV

CS771: Introduction to Machine Learning

Purushottam Kar

Recap of Last Lecture

Some nice continuous distributions

Uniform distribution – support is within an internal – no partiality within that!

Gaussian distribution – support is entire $\mathbb{R}$ – concentrates around the mean.

Concentrates strongly if variance small, weakly if variance large

Sum of two independent Gaussians is Gaussian, scaled Gaussian is a Gaussian

Tail Rule for Gaussians – deviation of more than $5\sigma < 0.000004$

Laplacian distribution – cousin of Gaussian – concentrates mode strongly

Probabilistic Regression

MLE with Gaussian Likelihood gives us least squares loss function

MLE with Laplacian Likelihood gives us absolute loss function

Probabilistic Regularization via Priors

Posterior, Maximum a Posteriori Estimator (MAP)

MAP can give us regularized or even constrained optimization problems

Be careful not to have strong priors (uninformed strong opinions are bad in life too ☺)

Random Vectors

Random vectors can be thought of as simply a collection of random variables arranged in an array $\mathbf{X} = [X_1, X_2, \dots, X_d]^\top$

No restriction on the random variables being independent or uncorrelated

PMF/PDF of $\mathbf{X}$ is simply the joint PMF/PDF of $\{X_1, X_2, \dots, X_d\}$

Can talk about marginal/conditional prob among $X_1, .. X_d$

Think of $X_1, X_2, \dots, X_d$ *as just a bunch of r.v.s*

$\mathbb{P}[X_2, X_3 \mid X_1, X_4, X_5]$

Since PMF/PDF of $\mathbf{X}$ is simply a joint PMF/PDF, all probability laws we learnt earlier continue to hold if we apply them correctly

Chain Rule, Sum Rule, Product Rule, Bayes Rule

Conditional/marginal variants of all these rules

Random Vectors

Expectation of a random variable is simply another vector (of same dim) of the expectations of the individual random variables

$$\mathbb{E}\mathbf{X} = [\mathbb{E}X_1, \mathbb{E}X_2, \ldots \mathbb{E}X_d]^\top$$

Linearity of expectation continues to hold: if $\mathbf{X}, \mathbf{Y}$ any two vector r.v. (not necessarily independent, then $\mathbb{E}[\mathbf{X} + \mathbf{Y}] = \mathbb{E}\mathbf{X} + \mathbb{E}\mathbf{Y}$

Scaling Rule: If $c \in \mathbb{R}$ is a constant then $\mathbb{E}[c \cdot \mathbf{X}] = c \cdot \mathbb{E}\mathbf{X}$

Dot Product Rule: If $\mathbf{a} \in \mathbb{R}^d$ is a constant vector, then $\mathbb{E}[\mathbf{a}^\top\mathbf{X}] = \mathbf{a}^\top\mathbb{E}\mathbf{X}$

Proof: $\mathbb{E}[\mathbf{a}^\top\mathbf{X}] = \mathbb{E}\left[\sum_{i=1}^d a_i X_i\right] = \sum_{i=1}^d \mathbb{E}[a_i X_i] = \sum_{i=1}^d a_i \cdot \mathbb{E}[X_i] = \mathbf{a}^\top\mathbb{E}\mathbf{X}$

Matrix Product Rule: If $A \in \mathbb{R}^{n\times d}$ is a constant matrix then

$$\mathbb{E}[A\mathbf{X}] = A\mathbb{E}\mathbf{X}$$

Proof: *Use Dot Product Rule n times*

Random Vectors

Mode easy to define: $\arg\max\limits_{X_1,\ldots,X_d} \mathbb{P}[X_1, \ldots, X_d]$

Median not easy to define – no unique definition

Definition 1: $\text{med}(\mathbf{X}) = [\text{med}(X_1)\,, \text{med}(X_2), \ldots, \text{med}(X_d)\,]^\top$

Definition 2: *minimizer of absolute distance (in this case L1 norm)*

$$\text{med}(\mathbf{X}) = \arg\min_{\mathbf{v}\in\mathbb{R}^d} \mathbb{E}[\|\mathbf{X} - \mathbf{v}\|_2]$$

Note: even here we still have $\mathbb{E}[\mathbf{X}] = \arg\min\limits_{\mathbf{v}\in\mathbb{R}^d} \mathbb{E}[\|\mathbf{X} - \mathbf{v}\|_2^2]$

Proof: $\mathbb{E}[\|\mathbf{X} - \mathbf{v}\|_2^2] = \mathbb{E}[\|\mathbf{X}\|_2^2] + \mathbb{E}[\|\mathbf{v}\|_2^2] - 2 \cdot \mathbf{v}^\top \mathbb{E}[\mathbf{X}]$

Taking derivative w.r.t $\mathbf{v}$ *and using first order optimality does the trick*

Random Vectors

Since random vectors are a bunch of real valued r.v.s, to specify the variance of this collection, need to have all pairwise covariances

Covariance

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \mathbb{V}X_1 & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathbb{V}X_1 & \dots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \dots & \mathbb{V}X_d \end{bmatrix}$$

Another cute formula

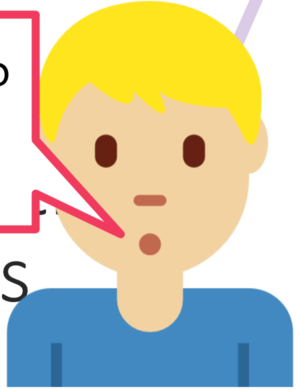
$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$, where $\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$

$\mathrm{Cov}(c \cdot \mathbf{X}) = c^2 \cdot \mathrm{Cov}(\mathbf{X})$

Random Vectors

Since random vectors are a bunch of … variance of this collection, need to have all pairwise covariances

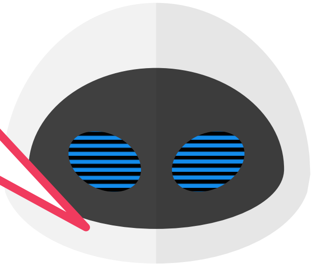
If $\mathbf{X}$ is a vector, isn't $\mathbf{X}\mathbf{X}^\top$ a matrix?
What does $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ even mean?



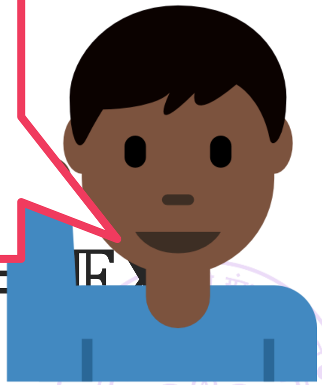
Covariance

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \mathbb{V}X_1 & \cdots \\ \mathrm{Cov}(X_2, X_1) & \cdots \end{bmatrix}$$

Just as a random vector is a collection of random variables arranged as a 1D array, a random matrix is a collection of r.v.s arranged as a 2D array!



Note that the (i,j)-th entry of the matrix $(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top$ is simply $(X_i-\mu_i)(X_j-\mu_j)$. Thus, (i,j)-th entry of $\mathbb{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top]$ is $\mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] = \mathrm{Cov}(X_i, X_j)$



Another cu…

$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top] = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$, where $\boldsymbol{\mu} = \mathbb{E}$…

$$\mathrm{Cov}(c \cdot \mathbf{X}) = c^2 \cdot \mathrm{Cov}(\mathbf{X})$$

Useful Operations on Vector R.V.

If $\mathbf{X} \in \mathbb{R}^m, \mathbf{Y} \in \mathbb{R}^n$ are two random vectors (not necessarily independent), then

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^\top] = \mathbb{E}[\mathbf{X}\mathbf{Y}^\top] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{Y}^\top \in \mathbb{R}^{m \times n}$$

where $\boldsymbol{\mu}_\mathbf{X} = \mathbb{E}\mathbf{X}$ and $\boldsymbol{\mu}_\mathbf{Y} = \mathbb{E}\mathbf{Y}$, $\text{Cov}(\mathbf{X}, \mathbf{Y})$

Dot Product Rule: If $\mathbf{a} \in \mathbb{R}^d$ is a constant vector, then $\mathbb{V}[\mathbf{a}^\top \mathbf{X}] = \mathbf{a}^\top \text{Cov}[\mathbf{X}]\mathbf{a}$

Proof: $\mathbb{V}[\mathbf{a}^\top \mathbf{X}] = \mathbb{E}[(\mathbf{a}^\top \mathbf{X})^2] - (\mathbf{a}^\top \boldsymbol{\mu}_\mathbf{X})^2 = \mathbb{E}[\mathbf{a}^\top \mathbf{X}\mathbf{X}^\top \mathbf{a}] - \mathbf{a}^\top \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top \mathbf{a}$
$= \mathbf{a}^\top \mathbb{E}[\mathbf{X}\mathbf{X}^\top]\mathbf{a} - \mathbf{a}^\top \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top \mathbf{a} = \mathbf{a}^\top \left(\mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\right)\mathbf{a} = \mathbf{a}^\top \text{Cov}[\mathbf{X}]\mathbf{a}$

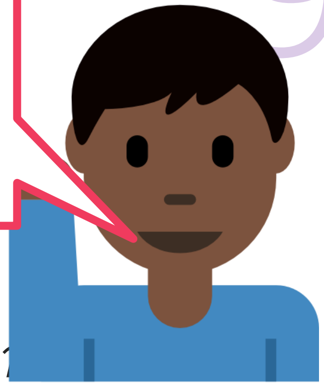
Matrix Product Rule: If $A \in \mathbb{R}^{n \times d}$ is a constant matrix then

$$\text{Cov}[A\mathbf{X}] = A\text{Cov}[\mathbf{X}]A^\top \in \mathbb{R}^{n \times n}$$

Proof: Try arguing similarly as the dot product rule

Useful Operations on V

Can you prove that the covariance matrix of any random vector is always a PSD matrix?



If $\mathbf{X} \in \mathbb{R}^m, \mathbf{Y} \in \mathbb{R}^n$ are two random

independent), then

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^\top] = \mathbb{E}[\mathbf{X}\mathbf{Y}^\top] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{Y}^\top \in \mathbb{R}$$

where $\boldsymbol{\mu}_\mathbf{X} = \mathbb{E}\mathbf{X}$ and $\boldsymbol{\mu}_\mathbf{Y} = \mathbb{E}\mathbf{Y}$, $\text{Cov}(\mathbf{X}, \mathbf{Y})$

Dot Product Rule: If $\mathbf{a} \in \mathbb{R}^d$ is a constant vector, then $\mathbb{V}[\mathbf{a}^\top\mathbf{X}] = \mathbf{a}^\top\text{Cov}[\mathbf{X}]\mathbf{a}$

Proof: $\mathbb{V}[\mathbf{a}^\top\mathbf{X}] = \mathbb{E}[(\mathbf{a}^\top\mathbf{X})^2] - (\mathbf{a}^\top\boldsymbol{\mu}_\mathbf{X})^2 = \mathbb{E}[\mathbf{a}^\top\mathbf{X}\mathbf{X}^\top\mathbf{a}] - \mathbf{a}^\top\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{a}$
$= \mathbf{a}^\top\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\mathbf{a} - \mathbf{a}^\top\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{a} = \mathbf{a}^\top\left(\mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\right)\mathbf{a} = \mathbf{a}^\top\text{Cov}[\mathbf{X}]\mathbf{a}$

Matrix Product Rule: If $A \in \mathbb{R}^{n \times d}$ is a constant matrix then

$$\text{Cov}[A\mathbf{X}] = A\text{Cov}[\mathbf{X}]A^\top \in \mathbb{R}^{n \times n}$$

Proof: Try arguing similarly as the dot product rule

Gaussian Random Vector

As in the scalar case, the *multivariate* Gaussian requires just the mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and the covariance $\Sigma \in \mathbb{R}^{d \times d}$ to be specified $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$\mathbb{P}[\mathbf{x} \mid \boldsymbol{\mu}, \Sigma] = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

Special case $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I_d$ called *standard Gaussian/Normal dist*

$$\mathbb{P}[\mathbf{x} \mid \mathbf{0}, I_d] = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\|\mathbf{x}\|_2^2\right) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right)$$

However, $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right)$ is simply $\mathcal{N}(0,1)$ i.e. we indeed have

$$\mathbb{P}[x_1, \dots, x_d \mid \mathbf{0}, \mathrm{I}] = \prod_{i=1}^d \mathbb{P}[x_i \mid 0,1]$$

All d coordinates of a standard Gaussian r.vec. are independent!