COMP3223: A quick review/introduction to probability theory

Srinandan Dasmahapatra

November 8, 2020

Reinterpret regression and classification probabilistically

- Softmax regression: Predict high probability of correct label c for data point x
  - For high $p(c|x)$ for $y_c = 1$, loss $-y_c \ln(w_c^\top x)$ low
  - Achieved by setting w_c large – overconfidence/overfitting
  - regularisation needed
- Linear regression: Predict output $\hat{y}$ given input x to make $r^2 = (y - \hat{y})^2$ small
  - Given family of functions $\hat{y} = f(x; w)$
  - lowering r^2 achieved by complex f with $\|w\|^2$ large
  - overfitting, fitting noise in data, regularisation needed
- Classification already in probabilistic language
- Interpret regression as finding model $f(\cdot; w)$ that makes large r^2 predictions improbable
- Regularisation by weight penalty viewed as imposing improbability of complex or large $\|w\|^2$ models even before data is seen

Outline: mostly about probability and statistics

- Basic probability theory and statistics
- Bivariate statistics (covariance) related to regression
- Bivariate continuous distributions
- Use linear dependence between two Gaussian random variables to motivate the form of the bivariate Gaussian distribution.

Basic definitions from probability theory: random variable, event/sample space

- The set of all events is Ω. Probability of event $A \subseteq \Omega$ is $P(A) \in [0, 1]$, $P(\Omega) = 1$.
- X variable, x value (specific event)
- **Probability mass function (pmf)** $P(X = x)$: quantifies how likely each possible outcome is:

$$P(X = x) = P_X(x) = P(x)$$
$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x)$$

- **Joint distribution** $P(A = a_i, B = b_j)$ is the probability that both events a_i and b_j occur.
- If events A, B **independent**, $P(A = a_i, B = b_j) = P(A = a_i)P(B = b_j)$: joint factorises into product of marginals
- **Conditional probability**, $P(A = a_i | B = b_j)$ is the probability that event a_i occurs given that event b_j has occurred: *information update*.

Bayes' rule for inference and inverse problems

- Given data $\mathbf{X}$ a set $\mathcal{H}$ of hypotheses $h_i \in \mathcal{H}$ that explains data.
- What is prob that given observation $\mathbf{X}$ was generated by some h_j?
- Equality of expressing joint in terms of conditionals:

$$P(A = a, B = b) = \begin{cases} P(B = b|A = a)P(A = a) \\ P(A = a|B = b)P(B = b) \end{cases}$$

- Leads to Bayes' rule:

$$P(B = b|A = a) = \frac{P(A = a|B = b)P(B = b)}{P(A = a)}.$$

- $P(\mathbf{X}|h_j)$ for each $h_j \in \mathcal{H}$ known; a **generative** mechanism: $h_j \to \mathbf{X}$
- Inverse problem: given data $\mathbf{X}$, find $P(h_i|\mathbf{X})$.

Expectation and variance characterise mean value of random variable and its dispersion.

- $x_i \sim P(X)$: $x_i, i = 1, \ldots, N$ random values drawn from $P(X)$
    - Example: die rolls, $x_i \in \{1, 2, 3, 4, 5, 6\}$
    - Example: individual i will infect x_i others, $x_i \in \{1, 2, 3, \ldots\}$.
- Collect data: $\mathcal{X} := \{x_1, x_2, \ldots, x_N\}$, mean of sample
- If $P(X)$ known, **Expectation**: $\mathbb{E}(X) = \sum_x P(X = x)x$ (population property) $\mathbb{E}(X) = \mathbb{E}X$ (notation).
- Expectation of a function of a random variable:

$$\mathbb{E}(f(X)) = \sum_x P(X = x)f(x)$$

- Moments= expectation of power of X: $M_k = \mathbb{E}X^k$
- Variance: Average (squared) fluctuation from the mean

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X - \mathbb{E}X)^2 && (1) \\ &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = M_2 - M_1^2 && (2) \end{aligned}$$

Bivariate distributions characterise systems of 2 observables.

- Joint distribution: $P(X = x, Y = y)$, a list of probabilities of all possible pairs of observations

- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(y=y)}$
- $X|Y$ has distribution $P(X|Y)$, a lookup-table of all possible $P(X = x|Y = y)$

Statistics of multivariate distributions:

- Conditional distributions are just distributions which have a (conditional) mean or variance.
- $\mathbb{E}(X|Y = y) = f(Y = y)$ fn of y "For each value of Y what is the average value of X?".
- $\mathbb{E}(X, Y) = \sum_{x,y} P(X = x, Y = y)(x, y) = (\mathbb{E}(X), \mathbb{E}(Y))$
- Covariance is the expected value of the product of fluctuations (deviations from mean):

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y \\ \mathrm{Var}(X) &= \mathrm{Cov}(X, X) \end{aligned}$$

- In finite sample, $\langle (X, Y) \rangle = (1/N) \sum_{n=1}^{N} (x_n, y_n)$
- Sample covariance $\sigma_{XY} = (1/N) \sum_{n=1}^{N} (x_n - \langle X \rangle)(y_n - \langle Y \rangle)$.
- Slope of regression line:

$$w_1 = \frac{\sigma_{XY}}{\sigma_{XX}}.$$

From linear regression - minimise $(\tilde{y}_n - w_1\tilde{x}_n)^2$

Fitting a straight line through points

Subtracting the average = data entering

- Subtract from each (x,y) the average:

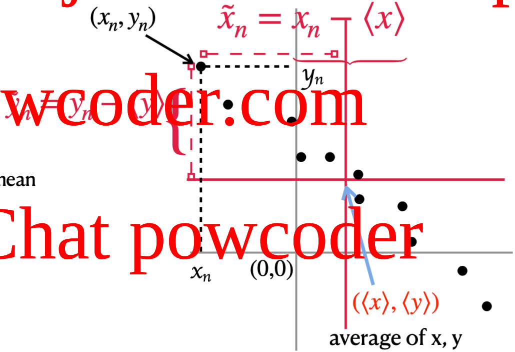
$$(\langle x\rangle, \langle y\rangle) = \frac{1}{N}\sum_n (x_n, y_n)$$

- The origin is shifted to the location of the mean
- Centred data: line goes through new origin

$$y_n = w_0 + w_1 x_n \Leftrightarrow \langle y\rangle = w_0 + w_1\langle x\rangle$$

- Subtract means:

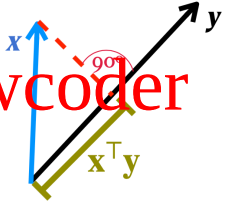
$$y_n - \langle y\rangle = w_1(x_n - \langle x\rangle) \Leftrightarrow \tilde{y}_n = w_1\tilde{x}_n$$



Exercise: differential calculus

Closed form solution to linear regression weights in terms of vector products

- $L(w) = (1/N)[(wx_1 - y_1)^2 + (wx_2 - y_2)^2 + \cdots + (wx_N - y_N)^2]$
- **Exercise**: In $L(w) = aw^2 + bw + c$, show
  - $a = (1/N)[x_1^2 + x_2^2 + \cdots + x_N^2]$
  - $b = (-2/N)[x_1 y_1 + x_2 y_2 + \cdots + x_N y_N]$
  - $0 = \left.\dfrac{\partial L(w)}{\partial w}\right|_{w=w^*} \implies w^* = -b/(2a) = \dfrac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}}$



Continuous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

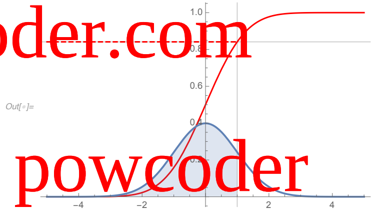
$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b p(x)dx \\ P(a \leq X \leq a + dx) &\approx p(a) \cdot dx \end{aligned}$$

- The cumulative distribution function is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.
- If A is an event, then

$$\begin{aligned} P(A) &= P(X \in A) = \int_{x \in A} p(x)dx \\ P(\Omega) &= P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1 \end{aligned}$$

Probability density function (pdf) and cumulative distribution function (cdf)

- $\text{CDF}(x) = \int_{-\infty}^{x} \text{PDF}(t)\,dt$
- Shaded area is value of the integral, $\text{CDF}(x = 1)$
- Red dashed line is value of the integral $\text{CDF}(x = 1)$



Continuous distributions: Mean, variance, conditionals have integrals, not sums

- Mean: $\mathbb{E}X = \int_x x \cdot p(x)dx$
- Variance: $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$
- Example: Uniform, Exponential
- If X has pdf $p(x)$, then $X|(X \in A)$ (restricted to domain A) has pdf

$$p_{X|A}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x)dx}$$

- Only makes sense if $P(A) > 0$!

Univariate Gaussian (Normal), $\mathcal{N}(\mu, \sigma)$

- Pdf of gaussian:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right)$$

- Statistics – population mean μ, variance σ^2:

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x p(x) dx = \mu \\ \mathbb{E}(X^2) &= \int_{-\infty}^{\infty} x^2 p(x) dx = \mu^2 + \sigma^2 \\ \mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \sigma^2 \end{aligned}$$

- Standard normal $\mathcal{N}(0, 1)$ has mean 0 and $\sigma = 1$: $p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

$$\int_{-\infty}^{\infty} p(z) dz = 1, \int_{-\infty}^{\infty} z p(z) dz = 0, \int_{-\infty}^{\infty} z^2 p(z) dz = 1.$$

Bivariate continuous distributions: Marginalisation, Conditioning and Independence

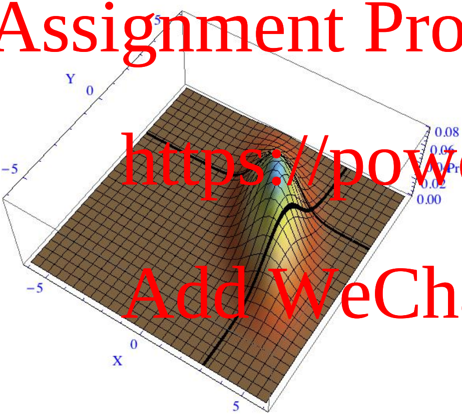
- $p_{X,Y}(x, y)$, joint probablity density function of X and Y
- $\int_x \int_y p(x, y) dx dy = 1$

- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y) dy$
- Conditional distribution:

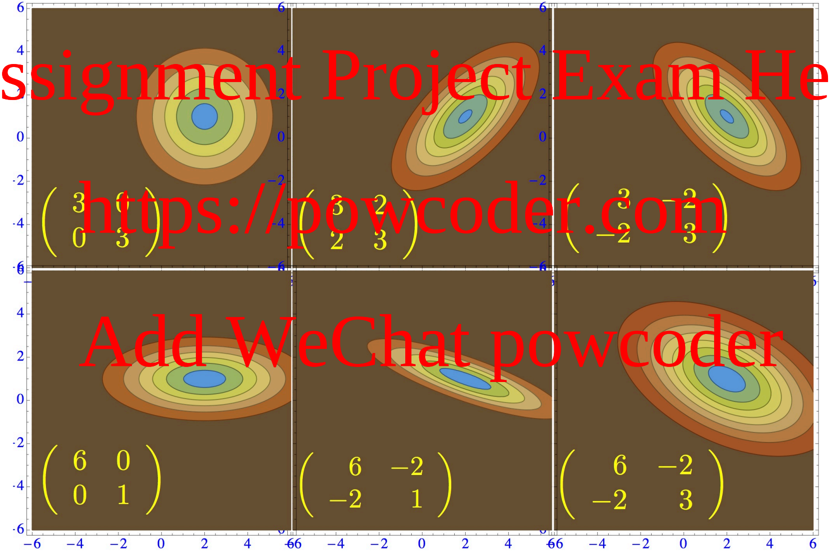
$$p(x|y)dx = \frac{p(x, y)dxdy}{p(y)dy}$$

- Independence: X and Y are independent if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$

Two dimensional Gaussian distributions



- The distribution on the left has mean $\mu = (2,1)^T$ and covariance matrix $\Sigma = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}$.
- The dark lines are for the conditional distributions $P(Y|X = 3.0)$ and $P(X|Y = 2.6)$. Notice that they are both Gaussian distributions.



$\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$
$\begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$
$\begin{pmatrix} 3 & -2 \\ -2 & 3 \end{pmatrix}$
$\begin{pmatrix} 6 & 0 \\ 0 & 1 \end{pmatrix}$
$\begin{pmatrix} 6 & -2 \\ -2 & 1 \end{pmatrix}$
$\begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix}$

Covariance matrix of X, Y linearly dependent Gaussian random variables

- Let $n_x \sim \mathcal{N}(0, \sigma_x)$, $n_y \sim \mathcal{N}(0, \sigma_y)$ two independent Gaussian r. v.
- Introduce 2 r. v.s $X = n_x$ and $Y = aX + n_y$, a real; $\mathbb{E}X = 0$, $\mathbb{E}Y = 0$.
- Compute components of covariance matrix $\boldsymbol{\Sigma}$ =Cov(X, Y): $\mathbb{E}X^2$, $\mathbb{E}XY$ and $\mathbb{E}Y^2$. Use Var$\mathcal{N}(0, \sigma) = \sigma^2$.
- $\mathbb{E}X^2 = \mathbb{E}n_x^2 = \sigma_x^2$

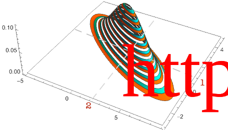
- $\mathbb{E}XY = \mathbb{E}n_x(an_x + n_y) = \mathbb{E}an_x^2 + \mathbb{E}n_x n_y = a\sigma_x^2$.
- $\mathbb{E}Y^2 = \mathbb{E}(a^2 n_x^2 + 2an_x n_y + n_y^2) = a^2\sigma_x^2 + \sigma_y^2$
- Assembling all terms for $\boldsymbol{\Sigma}$ and noting $\boldsymbol{\Sigma}^{-1}$ for reference:

$$\boldsymbol{\Sigma} = \begin{pmatrix} {\sigma_x}^2 & a{\sigma_x}^2 \\ a{\sigma_x}^2 & a^2{\sigma_x}^2 + {\sigma_y}^2 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\sigma_x^2} + \frac{a^2}{\sigma_y^2} & -\frac{a}{\sigma_y^2} \\ -\frac{a}{\sigma_y^2} & \frac{a^2}{\sigma_y^2} \end{pmatrix}$$

Example of 2-dimensional Gaussian distribution

- Given mean and covariance matrix of 2D Gaussian:

$$\mathcal{N}\left(x, y; \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{pmatrix} 6 & -2 \\ -2 & 1 \end{pmatrix}\right),$$



- compare cov. mat. with that of $Y = aX + n_y$, $X = n_x$

$$\mathbf{\Sigma} = \begin{pmatrix} {\sigma_x}^2 & a{\sigma_x}^2 \\ a{\sigma_x}^2 & a^2{\sigma_x}^2 + {\sigma_y}^2 \end{pmatrix} \implies \begin{cases} a = -1/3, \\ \sigma_x^2 = 6 \\ \sigma_y^2 = 1/3 \end{cases}$$

- Note negative slope, narrower distribution for y.
- How to set contour lines – lines of equal probability (equal height)?
- Express exponent in Gaussian as $e^{Q(x,y)}$
- Locus of pairs (x, y) so that $Q(x, y) =$ constant. (Called level sets.)

Obtain quadratic form $Q(x,y)$ from inverse covariance matrix for $X = n_x$, $Y = aX + n_y$

- Joint distribution (since n_x, n_y are independent)

$$p(X=x, Y=y) = \frac{1}{Z_x}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma_x}\right)^2\right)\frac{1}{Z_y}\exp\left(-\frac{1}{2}\left(\frac{y-ax}{\sigma_y}\right)^2\right)$$

- Consider the exponent in the joint distribution

$$-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{(y-ax)^2}{\sigma_y^2}\right) = -\frac{1}{2}\left((\frac{1}{\sigma_x^2}+\frac{a^2}{\sigma_y^2})x^2 - 2\frac{a}{\sigma_y^2}xy + \frac{1}{\sigma_y^2}y^2\right)$$

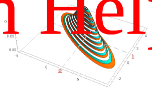
- The exponent $e^{Q(x,y)}$ has quadratic form $Q(x,y) = -\frac{1}{2}(x \;\; y)\Lambda\begin{pmatrix} x \\ y \end{pmatrix}$:

$$Q(x,y) = -\frac{1}{2}(x \;\; y)\begin{pmatrix} \frac{1}{\sigma_x^2}+\frac{a^2}{\sigma_y^2} & -\frac{a}{\sigma_y^2} \\ -\frac{a}{\sigma_y^2} & \frac{1}{\sigma_y^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}, \boldsymbol{\Lambda} \text{ turns out} = \boldsymbol{\Sigma}^{-1}.$$

Explicit form for 2-dimensional Gaussian distribution

To explicitly write the term in the exponent of

$$\mathcal{N}\left(x, y; \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 6 & -2 \\ -2 & 1 \end{bmatrix}\right)$$



ss $(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$ (precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$):

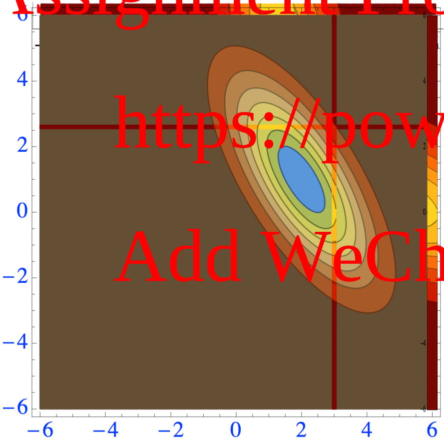
$$\underbrace{\frac{1}{Z}}_{\text{normalisation}} \exp\left(-\frac{1}{2\cdot 2}\begin{bmatrix} x-2, & y-1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 6 \end{bmatrix}\begin{bmatrix} x-2 \\ y-1 \end{bmatrix}\right),$$

where the inverse of the covariance matrix has been inserted and its determinant $= 2$ is in the denominator. This evaluates to

$$\left(-\frac{x^2}{4} - xy + 2x - \frac{3y^2}{2} + 5y - \frac{9}{2}\right).$$

The normalisation factor is $1/(2\sqrt{2}\pi)$.

Conditionals on contour plot



- These are contour plots of the (same) Gaussian pdf with mean $\mu = (2, 1)^{\mathsf{T}}$ and covariance matrix $\Sigma = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}$.
- Notice how the negative correlation between deviations from the means for X and Y.
- The light bands are for the conditional distributions $P(Y|X = 3.0)$ and $P(X|Y = 2.6)$. They are shown on the right and top, and they display a narrower distribution than the 2-d version.

General form for Gaussian distributions

A p-dimensional random variable $\mathbf{X} = (X_1, \ldots, X_p)$ has a probability density function $p(\mathbf{X}) \prod_i^p dX_i$ given by a *multivariate Gaussian distribution* specified by its mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$p(\mathbf{X}) = p(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right),$$

where $|\cdot|$ is the determinant. Equivalently, if $\mathbf{X} = \mathbf{x}$, a value taken by the random variable $\mathbf{X}$ distributed as a (multivariate) normal, we write

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Summary and looking ahead

- Recast task of reducing loss of model as that of reducing improbability of predictive model
- Formalise probabilistic modelling

- Relate statistics of distributions and samples to linear regression
- Interpreting 2-dimensional Gaussians
- Lab 3 and Chapter 2 of FCML addresses maximum likelihood estimation

- Next steps: regularisation to control increase of learned weight norms
- Bayesian: priors shape expectations of data modelling in absence of data (domain understanding)