COMP3223: A quick review/introduction to

November 8, 202

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

- Line $y - \hat{y})^2$
  smal

  - $\hat{y} = f(x; w)$
  - lowering r^2 achieved by complex f with
  - overfitting, fitting noise in data, regularisation r

- 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

- Basi
- Biva
- Biva
- Use linear dependence between two Gaussian ra
  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)
- **Pro** ... ach poss

$$P(X = x) = P_X(x) = P(x)$$
$$P(A) = P(x \in A) = \sum$$

- **Joint distribution** $P(A = a_i, B = b_j)$ ... 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:

- Leads to Bayes' rule:

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

- $P(\mathbf{X}|h_j)$ 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, $x_i \in \{1, 2, 3, \ldots\}$.
- Coll
- If $P($ tion prop

- Exp

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

- Moments= expectation of power of X: k
- Variance: Average (squared) fluctuation from the mean

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 \quad (1)$$

$$= \mathbb{E}X^2 - (\mathbb{E}X)^2 = M_2 - M_1^2 \quad (2)$$

Bivariate distributions characterise systems of 2 observables.

- Joint irs of obs

- Mar
- Conditional distribution: $P(X = x|Y = y$ $\frac{P(X=x,Y=y)}{}$
- $X|Y$ has distribution $P(X|Y)$, a lookup-table o $Y = y)$

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?".
- $\mathbb{E}(X$
- Cov
  from

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

- In finite sample, $\langle (X, Y) \rangle = (1/N) \sum_{n=1}^{N}$
- 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 $(\tilde{y}_n - w_1 \tilde{x}_n)^2$

Continuous random variables

- A random variable X is continuous if its sample space $\mathcal{X}$ is uncountable.

- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

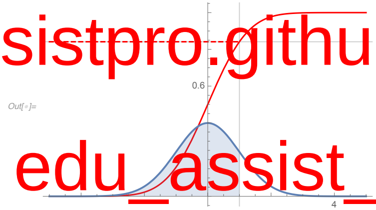
- The cumulative distribution function is F t
$p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)d$

- 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)

- CDF
- Shad
  CDF($x = 1$)
- Red dashed line is value of the integral CDF($x = 1$)



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

- Mean: $X = \quad x \; p(x)dx$
- Vari
- Exa

- If X has pdf $p(x)$, then $X|(X \in A)$ (restricted to domain A) has pdf

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

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

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

- Pdf of gaussian:

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

- Stati

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

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

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

Bivariate continuous distributions: Marginalisation, Conditioning and Independence

- $p_{X,}$ (x, y), joint probablity density function of X and Y
- $\int_x \int$

- Mar
- Conditional distribution:

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

- 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
  d covariance
  $-2 \Big)$.

- The dark lines are for the

$= 2.6)$.
an
distributions.

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; $\mathbb{E}X = 0$, $\mathbb{E}Y = 0$.
- Co … ^{2}, $\mathbb{E}XY$ and $\mathbb{E}Y^2$

- $\mathbb{E}X^2$ …
- $\mathbb{E}X$ … x x y x x y x
- $\mathbb{E}Y^2 = \mathbb{E}(a^2 n_x^2 + 2a n_x n_y + n_y^2) = a$ …
- Assembling all terms for $\mathbf{\Sigma}$ and noting $\mathbf{\Sigma}$ …

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

Example of 2-dimensional Gaussian distribution

- Given mean and covariance matrix of 2D Gaussian:

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

- n_y, $X = n_x$

$a = -1/3$,

$\Sigma =$ x x $a\sigma_x{}^2$ a^2

$\sigma_x^2 = 6$

$\sigma_y^2 = 1/3$

- 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)

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

- Con

$$-\frac{1}{2}\left(\frac{\cdots}{\sigma_x^2} + \frac{\cdots}{\sigma_y^2}\right) = -\frac{1}{2}\left(\left(\frac{\cdots}{\sigma} \cdots\right)^2 \cdots \frac{1}{\sigma_y^2}y^2\right)$$

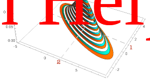
- The exponent $e^{Q(x,y)}$ has quadratic form $\cdots \begin{pmatrix} x \\ y \end{pmatrix}$:

$$Q(x, y) = -\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} \frac{1}{\sigma_x^2} + \frac{a^2}{\sigma_y^2} & -\frac{a}{\sigma_y^2} \\ -\frac{a}{\sigma_y^2} & \frac{1}{\sigma_y^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$, $\boldsymbol{\Lambda}$ 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 $(x - \mu$

$$\underbrace{\quad}_{\text{normalisation}} \; \ldots \Bigg]\Bigg),$$

where the inverse of the covariance matrix has been insert

$= 2$ 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
T
iance matrix

-

ns
Y

- 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 $\boldsymbol{\mu}$ a

$$p(\mathbf{X}) = \ldots \mathbf{X} - \boldsymbol{\mu}) \Big),$$

where $|\cdot|$ is the determinant. Equivalently, if $\mathbf{X}$ ... variable $\mathbf{X}$ distributed as a (multivariate) normal, we w

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

Summary and looking ahead

- Recast task of reducing loss of model as that of reducing improbability of predictive model
- For
- Rela

- Inter
- Lab 3 and Chapter 2 of FCML addresses maximum lik

- Next steps: regularisation to control increase of e
- Bayesian: priors shape expectations of data model (domain understanding)