Probabilistic ML

CS771: Introduction to Machine Learning

Purushottam Kar

Recap of Last Lecture

Creating events from random variables and vice versa: indicator r.v.

Union, intersection, complement of events, de Morgan’s Laws

Calculating probabilities of union/intersection/complement of events

Independence, conditional independence

Expectation, Mode, Median, (co)Variance: empirical counterparts

Rules of expectation: sum (linearity of expectation), product, LOTUS

Rules of (co)variance: sum, shift, scaling

Conditional Blah

The notation $[\cdot \mid \cdot]$ is used to express how one quantity behaves when some other quantities are fixed to some given values

These “other” quantities could be random variables themselves, or even constants. Sometimes we condition just to clarify exactly what those constants are

For example we could ask, what is the probability of me misclassifying a test data point $(\mathbf{x}, y) \sim \mathcal{D}$ *if I use a model* $\mathbf{w}$*?* $\mathbb{P}[y \cdot \mathbf{w}^\top \mathbf{x} < 0 \mid \mathbf{w}]$

Here $\mathbf{w}$ *is not a random variable (it could be in other settings but here it is not)*

We previously saw conditional probabilities $\mathbb{P}[X = 1 \mid Y = 2] = \frac{\mathbb{P}[X=1, Y=2]}{\mathbb{P}[Y=2]}$

Let us see other quantities that can be defined conditionally

Conditional Blah

Conditional Expectation $\mathbb{E}[X \mid Y = y_0] \triangleq \sum_{x \in S_X} x \cdot \mathbb{P}[X = x \mid Y = y_0]$

Conditional Variance $\mathbb{V}[X \mid Y = y_0] \triangleq \mathbb{E}[(X - \mu)^2 \mid Y = y_0]$ where we have $\mu = \mathbb{E}[X \mid Y = y_0]$

Conditional Covariance $\mathbf{Cov}[X, Y \mid Z = z_0]$

$= \mathbb{E}[(X - \mu_X) \cdot (Y - \mu_Y) \mid Z = z_0] = \mathbb{E}[XY \mid Z = z_0] - \mu_X \cdot \mu_Y$ where $\mu_X = \mathbb{E}[X \mid Z = z_0]$ and $\mu_Y = \mathbb{E}[Y \mid Z = z_0]$

Conditional Mode $\mathbf{mode}[X \mid Y = y_0] = \underset{x \in S_X}{\arg\max}\, \mathbb{P}[X = x \mid Y = y_0]$

Similarly we can define conditional median etc but not very popular

Note: these rules do not require X, Y, Z to be independent at all!!

Conditional Blah

Rules of expectation (sum, scaling, LOTUS, product) all continue to hold even with conditional except that all expectation are conditional

$\mathbb{E}[X + Y \mid Z = z_0] = \mathbb{E}[X \mid Z = z_0] + \mathbb{E}[Y \mid Z = z_0]$

$\mathbb{E}[c \cdot X \mid Z = z_0] = c \cdot \mathbb{E}[X \mid Z = z_0]$

$\mathbb{E}[g(X) \mid Z = z_0] = \sum_{x \in S_X} g(x) \cdot \mathbb{P}[X = x \mid Z = z_0]$

If $X \perp\!\!\!\perp Y \mid Z$ then $\mathbb{E}[X \cdot Y \mid Z = z_0] = \mathbb{E}[X \mid Z = z_0] \cdot \mathbb{E}[Y \mid Z = z_0]$

Rules of variance and covariance also continue to hold if we systematically condition all expressions involved in those rules

Note: conditioning must be the same everywhere, i.e. may happen that

$$\mathbb{E}[X + Y \mid Z = z_0] \neq \mathbb{E}[X \mid Z = z_1] + \mathbb{E}[Y \mid Z = z_2]$$

Probabilistic ML

Till now we have mostly ML techniques that assign a label for every data point (label ∈ ± 1 for binary classification, $[C]$ for multiclass classification, $\mathbb{R}$ for regression etc)

LwP, kNN, DT, linear models

Probabilistic ML techniques, given a data point, do not output a single label, they instead output a distribution over all possible labels

For binary classification, output a PMF over $\{-1,1\}$*, for multiclassification, output a PMF over* $\{1,2,\ldots,C\}$*, for regression … wait for another lecture*

The probability of a label in the output PMF indicates how likely does the ML model think that label is the correct one for that data point

***Warning**: a new (possibly different) PMF is output for every data point. The support of the PMF is all possible labels (even very unlikely ones are included)*

Probabilistic ML

Say we have somehow learnt a PML model $\mathbf{w}$ which, for a data point $\mathbf{x}$, gives us a PMF $\mathbb{P}[Y \mid \mathbf{x}, \mathbf{w}]$ over the set of all possible labels, say $\mathcal{Y}$

$\mathcal{Y} = \{-1, +1\}$ *for binary classification,* $\mathcal{Y} = [C]$ *for multiclassification*

Note that we conditioned on $\mathbf{x}, \mathbf{w}$ which are not r.v. at the moment but nevertheless fixed since we are looking at the *data point* $\mathbf{x}$ *using model* $\mathbf{w}$

We may use this PMF in very creative ways

Predict the mode of this PMF if someone wants a single label predicted

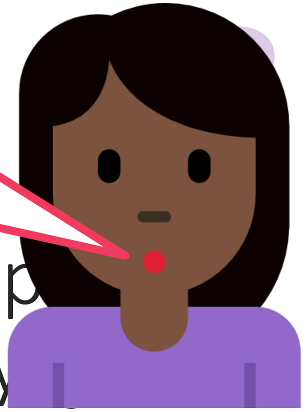
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \mathbb{P}[Y = y \mid \mathbf{x}, \mathbf{w}]$$

May use the median/mean as well – wait for a couple of lectures

Use $\mathbb{P}[Y = \hat{y} \mid \mathbf{x}, \mathbf{w}]$ *to find out if the ML model is confident about its prediction or totally confused about which label is the correct one!*

May use variance of $\mathbb{P}[Y \mid \mathbf{x}, \mathbf{w}]$ *to find this as well (low variance = very confident prediction and high variance = less confident/confused prediction)*

Exactly! Suppose we have three classes and for a data point, the ML model gives us the PMF $[0.3, 0.4, 0.3]$. The second class does win being the mode but the model seems not very certain about this prediction (only 40% confidence).

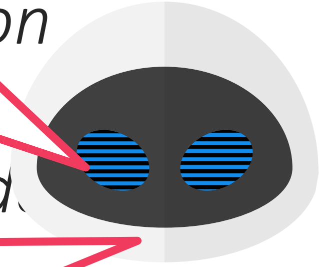


Say we have somehow learnt a PML model $\mathbf{w}$ which, for a data p

gives us a PMF $\mathbb{P}[Y \mid \mathbf{x}, \mathbf{w}]$ over the set of all possible labels, say

True! Suppose on another data point, the model gives us the PMF $[0.05, 0.1, 0.85]$. The third class wins being the mode and I extremely certain about this prediction (since I am giving a very high 85% confidence in this prediction).

on



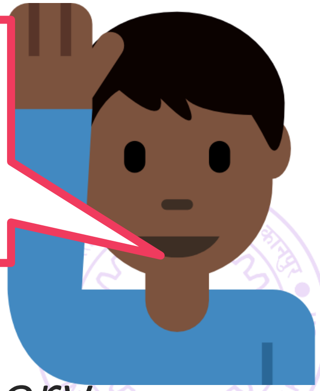
nevertheless fixed since we are looking at the *data point* $\mathbf{x}$ *using mod*

I could not agree more. However, in many ML applications (e.g. active learning) if we find that the model is making unsure predictions, we can switch to another model or just ask a human to step in. Thus, confidence info can be used fruitfully

ed

$$\hat{y} = \arg\max \mathbb{P}[Y = y \mid \mathbf{x}, \mathbf{w}]$$

Warning! Just because a prediction is made with more confidence does not mean it must be correct. It may happen that the 40% confidence prediction in the first case is correct but the high 85% confidence prediction in the second case is wrong!



prediction or totally confused about which label is the correct one!

May use variance of $\mathbb{P}[Y \mid \mathbf{x}, \mathbf{w}]$ *to find this as well (low variance = very confident prediction and high variance = less confident/confused prediction)*