Introduction to Bayesian Statistics

Richard Yi Da Xu

School of Computing & Communication, UTS

January 31, 2017

Random variables

- **Pre-university:** A number is just a fixed value.

When we talk about probabilities:

- When X is a continuous random variable, it has a probability density function (pdf)
- When X is a discrete random variable, it has a probability mass function (pmf)

$p(x) = p(X = x)$ means that:
The probability when a random variable X is equal to a fixed number x, i.e.,

the probablity that number of machine learning participants = 20

Mean or Expectation

- discrete case:

$$\mu = \mathbb{E}(X) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- continous case:

$$\mu = \mathbb{E}(X) = \int_{x \in \mathbb{S}} x p(x) \mathrm{d}x$$

- can also measure the expecation of a function:

$$\mathbb{E}(f(X)) = \int_{x \in \mathbb{S}} f(x) p(x) \mathrm{d}x$$

For example,

$$\mathbb{E}(\cos(X)) = \int_{x \in \mathbb{S}} \cos(x) p(x) \mathrm{d}x \qquad \mathbb{E}(X^2) = \int_{x \in \mathbb{S}} x^2 p(x) \mathrm{d}x$$

- What about $f(\mathbb{E}(X))$: Discuss later when we discuss Jensens Equality in Expecation-Maximization

Variances an intuitive explanation

- You have data $X = \{2, 3, 3, 2, 1, 4\}$, i.e., $x_1 = 2, x_2 = 3, \dots x_6 = 4$
- You have the mean:

$$\mu = \frac{2+3+3+2+1+4}{6} = 2.5$$

- The variance is then:

$$\mathbb{VAR}(data) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Division by N is intuitive. Otherwise, more data implies more variance
- Also think about what kind of values can $\mathbb{VAR}$ and σ take? - we will look at what kind of distribuiton is required for them.

Two alternative expression:

People sometimes use:

- You have data $X = \{2, 3, 3, 2, 1, 4\}$, i.e., $x_1 = 2, x_2 = 3, \dots x_6 = 4$

$$\mathbb{VAR}(X) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} 2x_i\mu + \frac{1}{N}\sum_{i=1}^{N}\mu^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\mu\underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\mu} + \mu^2$$

$$= \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu^2$$

It's easy to verify that both sides are the same

Other times, people use:

- You have data $X = \{1, 2, 3, 4\}$, and $P(X = 1) = \frac{2}{6}$, $P(X = 2) = \frac{2}{6}$, $P(X = 3) = \frac{1}{6}$ and $P(X = 4) = \frac{1}{6}$.

Discrete $: \mathbb{VAR}(X) = \sigma^2 = \sum_{x \in X}(x - \mu)^2 p(x)$

Continous $: \mathbb{VAR}(X) = \sigma^2 = \int_{x \in X}(x - \mu)^2 p(x)$

$$\mathbb{VAR}(X) = \sum_{x \in X}\left(x^2 - 2\mu x + \mu^2\right)p(x)$$

$$= \sum_{x \in X} x^2 p(x) - 2\sum_{x \in X}\mu x p(x) + \sum_{x \in X}\mu^2 p(x)$$

$$= \sum_{x \in X} x^2 p(x) - 2\mu\underbrace{\sum_{x \in X} x p(x)}_{\mu} + \mu^2\underbrace{\sum_{x \in X} p(x)}_{1}$$

$$= \left(\sum_{x \in X} x^2 p(x)\right) - \mu^2$$

Numerical example

First version

- $X = \{2, 3, 3, 2, 1, 4\}$, i.e., $x_1 = 2, x_2 = 3, \dots x_6 = 4$

$$\mathbb{VAR}(X) = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right) - \mu^2$$

$$= \frac{1}{6}(2 - 2.5)^2 + (3 - 2.5)^2 + (3 - 2.5)^2$$
$$+ (2 - 2.5)^2 + (1 - 2.5)^2 + (4 - 2.5)^2$$
$$\approx 0.917$$

Second version

- $X = \{1, 2, 3, 4\}$, and $P(X = 1) = \frac{2}{6}$, $P(X = 2) = \frac{2}{6}$, $P(X = 3) = \frac{1}{6}$ and $P(X = 4) = \frac{1}{6}$.

Discrete : $\mathbb{VAR}(X) = \sigma^2 = \sum_{x \in X} (x - \mu)^2 p(x)$

Continous : $\mathbb{VAR}(X) = \sigma^2 = \int_{x \in X} \underbrace{(x - \mu)^2}_{f(x)} p(x)$

$$\mathbb{VAR}(X) = \left(\sum_{x \in X} x^2 p(x) \right) - \mu^2$$

$$(1 - 2.5)^2 \frac{1}{6} + (2 - 2.5)^2 \frac{2}{6} + (3 - 2.5)^2 \frac{2}{6} +$$
$$(4 - 2.5)^2 \frac{1}{6}$$
$$\approx 0.917$$

Both sides are the same

Important fact of the Variances

$$
\begin{aligned}
\mathbb{VAR}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] &= \int_{x\in\mathbb{S}} (x-\mu)^2 p(x)\mathrm{d}x \\
&= \int_{x\in\mathbb{S}} x^2 p(x)\mathrm{d}x - 2\mu \int_{x\in\mathbb{S}} xp(x)\mathrm{d}x + \int_{x\in\mathbb{S}} \mu^2 xp(x)\mathrm{d}x \\
&= \mathbb{E}(\mathbf{X}^2) - (\mathbb{E}(\mathbf{X}))^2
\end{aligned}
$$

Think $\mathbb{VAR}(X)$ as "mean-subtracted" second order moment of random variable X.

Joint distributions

- The following is a tablet form of joint density $\Pr(X, Y)$:

	$Y=0$	$Y=1$	$Y=2$	Total
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
Total	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

- This table shows $\Pr(X, Y)$ or $\Pr(X = x, Y = y)$.
- For example, $p(X = 1, Y = 1) = \frac{6}{15}$:
- **exercise** what is the probablity that $X = 2, Y = 1$?
- **exercise** what is the probablity that $X = 3, Y = 2$?
- **exercise** what is the value of:

$$\sum_{i=0}^{2} \sum_{j=0}^{2} \Pr(X = i, Y = j)?$$

Marginal distributions

	$Y=0$	$Y=1$	$Y=2$	Total
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
Total	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

- Using **sum rule**, the **marginal distribution** tells us that:

$$\Pr(X) = \sum_{y \in \mathbb{S}_y} \Pr(x, y) \qquad \text{or} \qquad p(X) = \int_{y \in \mathbb{S}_y} p(x, y) \mathrm{d}y$$

- For example:

$$\Pr(Y=1) = \sum_{i=0}^{2} \sum_{j=0}^{2} p(x=i, y=1) = \frac{3}{15} + \frac{6}{15} + \frac{0}{15} = \frac{9}{15}$$

- **exercise** what is $\Pr(X=2)$ and $\Pr(X=1)$?

Conditional distributions

	$Y=0$	$Y=1$	$Y=2$	Total
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
Total	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

- Conditional density:

$$p(X|Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(Y|X)p(X)}{p(Y)} = \frac{p(Y|X)p(X)}{\sum_X p(Y|X)p(X)}$$

- What about $p(X|Y=y)$? Pick an example:

$$p(X=1|Y=1) = \frac{p(X=1,Y=1)}{p(Y=1)} = \frac{6/15}{9/15} = \frac{2}{3}$$

Conditional distributions: Exercise

	$Y=0$	$Y=1$	$Y=2$	Total
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
Total	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

- The formulation for conditional density:

$$p(X|Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(Y|X)p(X)}{p(Y)} = \frac{p(Y|X)p(X)}{\sum_X p(Y|X)p(X)}$$

- **exercise**: What is $p(X=2|Y=1)$?
- **exercise**: What is $p(X=1|Y=2)$?

Independence

If X and Y are independent:

- $p(X|Y) = p(X)$
- $p(X, Y) = P(X)P(Y)$
- Both factors are related when A and B are independent:

$$p(X|Y) = \frac{p(X, Y)}{p(Y)} = \frac{p(X)p(Y)}{p(Y)} = p(X)$$

	$Y = 0$	$Y = 1$	$Y = 2$	Total
$X = 0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X = 1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X = 2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
Total	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

X and Y are NOT independent

	$Y = 0$	$Y = 1$	$Y = 2$	Total
$X = 0$	$\frac{18}{225}$	$\frac{54}{225}$	$\frac{18}{225}$	$\frac{6}{15}$
$X = 1$	$\frac{24}{225}$	$\frac{72}{225}$	$\frac{24}{225}$	$\frac{8}{15}$
$X = 2$	$\frac{3}{225}$	$\frac{9}{225}$	$\frac{3}{225}$	$\frac{1}{15}$
Total	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

X and Y are independent

Conditional Independence

- Imagine we have three random variables: X, Y and Z:
- Once we know Z, then knowing Y does NOT tell us any additional information about X
- Therefore:

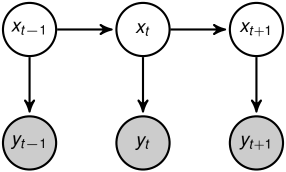
$$\Pr(X|Y,Z) = \Pr(X|Z)$$

- This means that X is conditionally independent of Y given Z.
- If $\Pr(X|Y,Z) = \Pr(X|Z)$, then what about $\Pr(X, Y|Z)$?

$$\begin{aligned}
\Pr(X, Y|Z) = \frac{\Pr(X, Y, Z)}{\Pr(Z)} &= \frac{\Pr(X|Y,Z)\Pr(Y,Z)}{\Pr(Z)} \\
&= \Pr(X|Y,Z)\Pr(Y|Z) \\
&= \Pr(X|Z)\Pr(Y|Z)
\end{aligned}$$

An example of Conditional Independence

We will study **Dynamic model** later.



From this model, we can see:

$$
\begin{aligned}
p(x_t|x_1,\dots,x_{t-1},y_1,\dots,y_{t-1}) &= p(x_t|x_{t-1}) \\
p(y_t|x_1,\dots,x_{t-1},x_t,y_1,\dots,y_{t-1}) &= p(y_t|x_t)
\end{aligned}
$$

Right now, think of if a given variable is the only item that “blocks” the path between two (or more) variables.

Another Example: Bayesian Linear Regression

We have data pairs:

- Input: $X = x_1, \dots x_N$
- Output: $Y = y_1, \dots y_N$

Each pair, x_i and y_i are related through model equation:

$$y_i = f(x_i|w) + \mathcal{N}(0, \sigma^2)$$

- Input alone isn't going to tell you model parameter: $p(w|X) = p(w)$
- Output alone isn't going to tell you model parameter: $p(w|Y) = p(w)$
- Obviously: $p(w|X, Y) \neq p(w)$

Posterior over parameter w:

$$p(w|x, y) = \frac{p(y|w, x)p(w|x)p(x)}{p(y|x)p(x)} = \frac{p(y|w, x)p(w)}{p(y|x)} = \frac{p(y|w, x)p(w)}{\int_w p(y|x, w)p(w)}$$

Expectation of Joint probabilities

Given that X, Y is a two-dimensional random variable:

- Continous case:

$$\mathbb{E}[f(X,Y)] = \int_{y\in\mathbb{S}_y}\int_{x\in S_x} f(x,y)p(x,y)\mathrm{d}x\mathrm{d}y$$

- Discrete case:

$$\mathbb{E}[f(X,Y)] = \sum_{i=1}^{N_i}\sum_{j=1}^{N_j} f(X=i, Y=j)p(X=i, Y=j)$$

Numerical Example:

	$Y=1$	$Y=2$	$Y=3$
$X=1$	0	$\frac{3}{15}$	$\frac{3}{15}$
$X=2$	$\frac{2}{15}$	$\frac{6}{15}$	0
$X=3$	$\frac{1}{15}$	0	0

p (X,Y)

	$Y=1$	$Y=2$	$Y=3$
$X=1$	6	7	8
$X=2$	3	6	2
$X=3$	1	8	6

f (X,Y)

$$\begin{aligned}\mathbb{E}[f(X,Y)] &= \sum_{i=1}^{N_i}\sum_{j=1}^{N_j} f(X=i,Y=j)p(X=i,Y=j) \\ &= 6\times 0 + 7\times\frac{3}{15} + 8\times\frac{3}{15} + 3\times\frac{2}{15} + 6\times\frac{6}{15} \\ &\quad + 2\times 0 + 1\times\frac{1}{15} + 8\times 0 + 6\times 0\end{aligned}$$

It's a useful property for later

$$
\begin{aligned}
\mathbb{E}(Y) &= \int_X \mathbb{E}(Y|X)p(X)\mathrm{d}x \\
&= \int_X \underbrace{\int_Y yp(Y|X)\mathrm{d}y}\, p(X)\mathrm{d}x = \int_X \int_Y yp(Y,X)\mathrm{d}y\mathrm{d}x \\
&= \int_Y y \left(\int_X p(Y,X)\mathrm{d}x \right) \mathrm{d}y \\
&= \int_Y yp(Y)\mathrm{d}y = \mathbb{E}(Y)
\end{aligned}
$$

Bayesian Predictive distribution

Put marginal distribution and Conditional Independence into a test:

- Very often, in machine learning, you want to compute the probability of new data y^* given training data Y, i.e., $p(y^*|Y)$. You assume there are some model explains both Y and y^*. The model parameter is θ:

$$p(y^*|Y) = \int_\theta p(y^*|\theta)p(\theta|Y)\mathrm{d}\theta$$

- Excercise, tell me why the above works?

Revisit Bayes Theorem

Instead of using arbitrary random variable symbols, we now use:

- θ for model parameter
- and $X = x_1, \ldots x_n$ for dataset:

$$\underbrace{p(\theta|X)}_{\text{posterior}} = \frac{\overbrace{\underbrace{p(X|\theta)}_{\text{likelihood}}\ \underbrace{p(\theta)}_{\text{prior}}}}{\underbrace{p(X)}_{\text{normalization constant}}} = \frac{p(X|\theta)p(\theta)}{\int_\theta p(X|\theta)p(\theta)}$$

An Intrusion Detection System (IDS) Example

The setting: Imagine out of all the TCP connections (say millons), 1% of which are intrusions:

- When there is an intrusion, the probability of system sends alarm is 87%.
- When there is no intrusion, the probability of system sends alarm is 6%.

- **Prior probability**:
  1% of which are intrusions
  $\implies p(\theta = \text{intrusion}) = 0.01 \qquad p(\theta = \text{no intrusion}) = 0.99$
- **Likelihood probability**:
  - given intrusion occur, probability of system sends alarm is 87%
  $$p(X = \text{alarm}|\theta = \text{intrusion}) = 0.87 \qquad p(X = \text{no alarm}|\theta = \text{intrusion}) = 0.13$$
  - given there is no intrusion, the probability of system sends alarm is 6%:
  $$p(X = \text{alarm}|\theta = \text{no intrusion}) = 0.06 \qquad p(X = \text{no alarm}|\theta = \text{no intrusion}) = 0.94$$

Posterior

- We are interested in **posterior probability**: $\Pr(\theta|X)$:
- There 2 two possible values for parameter θ and 2 possible observation X
- Therefore, there are 4 **rates** we need to compute:
  - **True Positive** When system sends alarm, probability of an intrusion occurs:
  $$\Pr(\theta = \text{intrusion}|X = \text{alarm})$$
  - **False Positive** When system sends alarm, probability that there is no intrusion:
  $$\Pr(\theta = \text{no intrusion}|X = \text{alarm})$$
  - **True Negative** When system sends no alarm, probability that there is no intrusion:
  $$\Pr(\theta = \text{no intrusion}|X = \text{no alarm})$$
  - **False Negative** When system sends no alarm, probability that an intrusion occurs:
  $$\Pr(\theta = \text{intrusion}|X = \text{no alarm})$$
- **Question** which are the two probabilities you'd like to maximise?

Apply Bayes Theorem in this setting

$$
\begin{aligned}
\Pr(\theta|X) &= \frac{\Pr(X|\theta)\Pr(\theta)}{\sum_\theta \Pr(X|\theta)\Pr(\theta)} \\
&= \frac{\Pr(X|\theta)\Pr(\theta)}{\Pr(X|\theta = \text{Intrusion})\Pr(\theta = \text{Intrusion}) + \Pr(X|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion})}
\end{aligned}
$$

Apply Bayes Theorem in this setting

True Positive rate When system sends alarm, what is the probability of an intrusion occurs:

$$
\begin{aligned}
&\Pr(\theta = \text{intrusion} | X = \text{alarm}) \\
&= \frac{\Pr(X = \text{alarm} | \theta = \text{intrusion}) \Pr(\theta = \text{intrusion})}{\Pr(X = \text{alarm} | \theta = \text{Intrusion}) \Pr(\theta = \text{Intrusion}) + \Pr(X = \text{alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{Intrusion})} \\
&= \frac{0.87 \times 0.01}{0.87 \times 0.01 + 0.06 \times 0.99} = 0.1278
\end{aligned}
$$

False Positive rate When system sends alarm, what is the probability that there is no intrusion:

$$
\begin{aligned}
&\Pr(\theta = \text{no intrusion} | X = \text{alarm}) \\
&= \frac{\Pr(X = \text{alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{no intrusion})}{\Pr(X = \text{alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{no intrusion}) + \Pr(X = \text{alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{no intrusion})} \\
&= \frac{0.06 \times 0.99}{0.87 \times 0.01 + 0.06 \times 0.99} = 0.8722
\end{aligned}
$$

Apply Bayes Theorem in this setting

False Negative When system sends no alarm, what is the probability that an intrusion occurs?

$$
\begin{aligned}
&\Pr(\theta = \text{intrusion}|X = \text{no alarm}) \\
&= \frac{\Pr(X = \text{no alarm}|\theta = \text{intrusion})p(\theta = \text{intrusion})}{\Pr(X = \text{no alarm}|\theta = \text{Intrusion})\Pr(\theta = \text{Intrusion}) + \Pr(X = \text{no alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no Intrusion})} \\
&= \frac{0.13 \times 0.01}{0.13 \times 0.01 + 0.94 \times 0.99} = 0.0014
\end{aligned}
$$

True Negative When system sends no alarm, what is the probability that there is no intrusion?

$$
\begin{aligned}
&\Pr(\theta = \text{no intrusion}|X = \text{no alarm}) \\
&= \frac{\Pr(X = \text{no alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion})}{\Pr(X = \text{no alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion}) + \Pr(X = \text{no alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion})} \\
&= \frac{0.94 \times 0.99}{0.87 \times 0.001 + 0.06 \times 0.99} = 0.9986
\end{aligned}
$$

Statistics way to think about Posterior Inference

The posterior inference is to find the best $q(\theta)$ to approximate $p(\theta|X)$, such that:

$$
\begin{aligned}
&\inf_{q(\theta)\in\mathcal{Q}} \left\{\mathrm{KL}(q(\theta)\|p(\theta)) - \mathbb{E}_{\theta\sim q(\theta)} \ln(p(X|\theta)\right\} \\
=&\inf_{q(\theta)\in\mathcal{Q}} \left\{\int_\theta \ln \frac{q(\theta)}{p(\theta)} q(\theta) - \int_\theta \ln(p(X|\theta)q(\theta)\right\} \\
=&\inf_{q(\theta)\in\mathcal{Q}} \left\{\int_\theta \left[\ln q(\theta) - (\ln p(\theta) + \ln p(X|\theta))\right] q(\theta)\right\} \\
=&\inf_{q(\theta)\in\mathcal{Q}} \left\{\int_\theta \left[\ln \frac{q(\theta)}{p(\theta)p(X|\theta)}\right] q(\theta)\right\} \\
=&\frac{1}{p(X)} \inf_{q(\theta)\in\mathcal{Q}} \left\{\int_\theta \left[\ln \frac{q(\theta)}{p(\theta|X)}\right] q(\theta)\right\} \\
=&\inf_{q(\theta)\in\mathcal{Q}} \left\{\mathrm{KL}(q(\theta)\|p(\theta|X))\right\}
\end{aligned}
$$