SFDS/DAE

University of the Witwatersrand

2025

Review Question

- Compute $\binom{10}{4}$

Revision Problem

- A coin is flipped five times and comes up heads all five times. Perform a two sided hypothesis test to see if the coin is fair.

Lesson Plan

- Review Question

Lesson Plan

- Review Question
- Review of Beta as a conjugate prior

Lesson Plan

- Review Question
- Review of Beta as a conjugate prior
- Some other conjugate prior distributions

Lesson Plan

- Review Question
- Review of Beta as a conjugate prior
- Some other conjugate prior distributions
- Multivariate distributions Discrete

Lesson Plan

- Review Question
- Review of Beta as a conjugate prior
- Some other conjugate prior distributions
- Multivariate distributions Discrete
- Multivariate distributions Continuous

Lesson Plan

- Review Question
- Review of Beta as a conjugate prior
- Some other conjugate prior distributions
- Multivariate distributions Discrete
- Multivariate distributions Continuous
- Riddle

Review Question

- Compute $\binom{10}{4}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times 9\times 8\times 7}{4\times 3\times 2\times 1}$
- Compute $\frac{10\times 9\times 8\times 7}{4\times 3\times 2}$
- Compute $\frac{10\times 9\times 7}{3}$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2\times1}$
- Compute $\frac{10\times9\times8\times7}{4\times3\times2}$
- Compute $\frac{10\times9\times7}{3}$
- Compute $10 \times 3 \times 7$

Review Question

- Compute $\binom{10}{4}$
- Compute $\frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1}$
- Compute $\frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2}$
- Compute $\frac{10 \times 9 \times 7}{3}$
- Compute $10 \times 3 \times 7$
- 210

The Beta as a conjugate prior

- We have a coin which has some unknown probability p of coming up heads. That is it's associated to a bernoulli distribution.

The Beta as a conjugate prior

- We have a coin which has some unknown probability *p* of coming up heads. That is it's associated to a bernoulli distribution.
- We start out thinking that *p* is distributed $Beta(\alpha, \beta)$. Commonly $Beta(1, 1)$ which is the uniform.

The Beta as a conjugate prior

- We have a coin which has some unknown probability p of coming up heads. That is it's associated to a bernoulli distribution.
- We start out thinking that p is distributed $Beta(\alpha, \beta)$. Commonly $Beta(1, 1)$ which is the uniform.
- Then some data comes in. That is we flip the coin n times and get k heads.

The Beta as a conjugate prior

- We have a coin which has some unknown probability p of coming up heads. That is it's associated to a bernoulli distribution.
- We start out thinking that p is distributed $Beta(\alpha, \beta)$. Commonly $Beta(1, 1)$ which is the uniform.
- Then some data comes in. That is we flip the coin n times and get k heads.
- Out posterior distribution on p is $Beta(\alpha + n - k, \beta + k)$.

The Beta as a conjugate prior

- We have a coin which has some unknown probability p of coming up heads. That is it's associated to a bernoulli distribution.
- We start out thinking that p is distributed $Beta(\alpha, \beta)$. Commonly $Beta(1, 1)$ which is the uniform.
- Then some data comes in. That is we flip the coin n times and get k heads.
- Out posterior distribution on p is $Beta(\alpha + n - k, \beta + k)$.
- This is because the Beta distribution is a conjugate prior to the bernoulli.

Other conjugate Priors

Likelihood	**Prior (Conjugate)**	**Posterior**
$X \sim \text{Binomial}(n, \theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \mid X \sim \text{Beta}(\alpha + x, \beta + n - x)$
$X \sim \text{Poisson}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \mid X \sim \text{Gamma}(\alpha + x, \beta + 1)$
$X \sim \mathcal{N}(\mu, \sigma^2)$ (known σ^2)	$\mu \sim \mathcal{N}(\mu_0, \tau^2)$	$\mu \mid X \sim \mathcal{N}(\mu_n, \tau_n^2)$ $$\tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$ $$\mu_n = \tau_n^2 \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)$$

Other conjugate Priors

Likelihood	**Prior (Conjugate)**	**Posterior**
$X \sim \text{Exponential}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \mid X \sim \text{Gamma}(\alpha + n, \beta + \sum x_i)$
$\mathbf{X} \sim \text{Multinomial}(n, \boldsymbol{\theta})$	$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$	$\boldsymbol{\theta} \mid \mathbf{X} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{X})$

Multivariate Distribution

- A **multivariate discrete distribution** describes a vector of random variables $(X_1, \ldots, X_d)$ taking values jointly.

Multivariate Distribution

- A **multivariate discrete distribution** describes a vector of random variables $(X_1, \ldots, X_d)$ taking values jointly.
- For example the probability mass function (PMF) of two zero-one random variables could be given as:

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.3	0.4
$X_1 = 1$	0.2	0.1

Multivariate Distribution

- A **multivariate discrete distribution** describes a vector of random variables $(X_1, \ldots, X_d)$ taking values jointly.
- For example the probability mass function (PMF) of two zero-one random variables could be given as:

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.3	0.4
$X_1 = 1$	0.2	0.1

- Or more generally by:

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	p_{00}	p_{01}
$X_1 = 1$	p_{10}	p_{11}

Multivariate Distribution

- It doesn't need to be binomial of course

Multivariate Distribution

- It doesn't need to be binomial of course

Y \ X	1	2	3	4	5
4	0.2	0	0	0	0.2
3	0	0.12	0	0.12	0
2	0.12	0	0.12	0	0
1	0	0	0	0	0.12

Multivariate Distribution

- One of the things we may wish to do here is to calculate the distribution of X or Y. This is called the **Marginal Distribution** of X or Y.

Multivariate Distribution

- One of the things we may wish to do here is to calculate the distribution of X or Y. This is called the **Marginal Distribution** of X or Y.
- This is done by summing probabilities for specific values. In the above table.

Multivariate Distribution

- One of the things we may wish to do here is to calculate the distribution of X or Y. This is called the **Marginal Distribution** of X or Y.
- This is done by summing probabilities for specific values. In the above table.
-

$$P(X = x) = \begin{cases} 0.32, & x = 1, \\ 0.12, & x = 2, \\ 0.12, & x = 3, \\ 0.12, & x = 4, \\ 0.32, & x = 5, \\ 0, & \text{otherwise.} \end{cases}$$

Multivariate Distribution

- Similarly for Y.

Multivariate Distribution

- Similarly for Y.
- $$P(Y=y)=\begin{cases}0.12, & y=1,\\ 0.24, & y=2,\\ 0.24, & y=3,\\ 0.4, & y=4,\\ 0, & \text{otherwise.}\end{cases}$$

Covariance

- $cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$

Covariance

- $cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$
- Turns out to be equal to $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Covariance

- $cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$
- Turns out to be equal to $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- $In our example \mathbb{E}[XY] = 2 * 0.12 + 4 * 0.2 + 6 * 0.12 + 6 * 0.12 + 12 * 0.12 + 5 * 0.12 + 20 * 0.2 = 8.52$

Covariance

- $cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$
- Turns out to be equal to $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- $In our example \mathbb{E}[XY] = 2 * 0.12 + 4 * 0.2 + 6 * 0.12 + 6 * 0.12 + 12 * 0.12 + 5 * 0.12 + 20 * 0.2 = 8.52$
- $\mathbb{E}[X] = 3$ and $\mathbb{E}[Y] = 2.92$

Covariance

- $cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$
- Turns out to be equal to $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- $In our example \mathbb{E}[XY] = 2 * 0.12 + 4 * 0.2 + 6 * 0.12 + 6 * 0.12 + 12 * 0.12 + 5 * 0.12 + 20 * 0.2 = 8.52$
- $\mathbb{E}[X] = 3$ and $\mathbb{E}[Y] = 2.92$
- $Cov(X, Y) = 8.52 - 3 \times 2.92 = -0.24$

Covariance

- $cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$
- Turns out to be equal to $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- $Inourexample \mathbb{E}[XY] = 2 * 0.12 + 4 * 0.2 + 6 * 0.12 + 6 * 0.12 + 12 * 0.12 + 5 * 0.12 + 20 * 0.2 = 8.52$
- $\mathbb{E}[X] = 3$ and $\mathbb{E}[Y] = 2.92$
- $Cov(X, Y) = 8.52 - 3 \times 2.92 = -0.24$
- Note that $Cor(X, Y) = \frac{Cov(X,Y)}{\sqrt{(V(X)V(Y)}}$ and is between -1 and 1

Multivariate Distribution

- We also care about conditional distributions.
- That is the distribution of Y given that X takes on a particular value or visa versa. For example
- $$P(Y = y|X = 1) = \begin{cases} 0.625, & y = 2, \\ 0.375, & y = 4, \\ 0, & \text{otherwise.} \end{cases}$$

Multivariate Distribution

- Consider the distribution:

$$f_{X,Y}(x,y) = \begin{cases} x+y, & 0 \leq x \leq 1, \ 0 \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Multivariate Distribution

- Consider the distribution:

$$f_{X,Y}(x,y) = \begin{cases} x+y, & 0 \leq x \leq 1, \;\; 0 \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

- Let's check that it is a distribution first!

Multivariate Distribution

- Consider the distribution:

$$f_{X,Y}(x,y) = \begin{cases} x+y, & 0 \le x \le 1, \ \ 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

- Let's check that it is a distribution first!
- It's always positive.

Multivariate Distribution

- $$\begin{aligned}\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dx\,dy &= \int_0^1 \int_0^1 (x+y)\,dx\,dy \\ &= \int_0^1 \Big[\underbrace{\int_0^1 x\,dx}_{\frac{1}{2}} + \underbrace{\int_0^1 y\,dx}_{y}\Big]\,dy \\ &= \int_0^1 \left(\tfrac{1}{2}+y\right)dy \\ &= \left[\tfrac{1}{2}y + \tfrac{1}{2}y^2\right]_0^1 = \tfrac{1}{2}+\tfrac{1}{2} = 1.\end{aligned}$$

Multivariate Distribution

- $$\begin{aligned}\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dx\,dy &= \int_0^1 \int_0^1 (x+y)\,dx\,dy \\ &= \int_0^1 \Big[\underbrace{\int_0^1 x\,dx}_{\frac{1}{2}} + \underbrace{\int_0^1 y\,dx}_{y}\Big]\,dy \\ &= \int_0^1 \left(\tfrac{1}{2} + y\right) dy \\ &= \left[\tfrac{1}{2}y + \tfrac{1}{2}y^2\right]_0^1 = \tfrac{1}{2} + \tfrac{1}{2} = 1.\end{aligned}$$
- It is indeed a distribution!

Marginal Distribution of X

- Once again, we might be interested in the dstrbution of X.

Marginal Distribution of X

- Once again, we might be interested in the dstrbution of X.
- To get this we integrate out Y.

Marginal Distribution of X

- Once again, we might be interested in the dstrbution of X.
- To get this we integrate out Y.
- $$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\ &= \int_0^1 (x+y)\,dy \\ &= \left[xy + \frac{1}{2}y^2 \right]_0^1 \\ &= x + \frac{1}{2}, \end{aligned}$$

Marginal Distribution of X

- Once again, we might be interested in the dstrbution of X.
- To get this we integrate out Y.
- $$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\ &= \int_0^1 (x+y)\,dy \\ &= \left[xy + \frac{1}{2}y^2 \right]_0^1 \\ &= x + \frac{1}{2}, \end{aligned}$$
- So
$$f_X(x) = \begin{cases} x + \frac{1}{2}, & 0 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Conditional Distributions

- $$f_{Y|X}(y \mid X = 0.5) = \frac{f_{X,Y}(0.5, y)}{f_X(0.5)} = \frac{0.5 + y}{0.5 + \frac{1}{2}} = 0.5 + y,$$

  $$0 \leq y \leq 1,$$

  $$f_{Y|X}(y \mid X = 0.5) = 0 \quad \text{otherwise.}$$

Conditional Distributions

- $$f_{Y|X}(y \mid X=0.5) = \frac{f_{X,Y}(0.5,y)}{f_X(0.5)} = \frac{0.5+y}{0.5+\frac{1}{2}} = 0.5+y,$$

  $0 \le y \le 1,$

  $$f_{Y|X}(y \mid X=0.5) = 0 \quad \text{otherwise.}$$

- $$f_{Y|X}(y \mid X=0.7) = \frac{f_{X,Y}(0.7,y)}{f_X(0.7)} = \frac{0.7+y}{0.7+\frac{1}{2}} = \frac{0.7+y}{1.2},$$

  $0 \le y \le 1,$

  $$f_{Y|X}(y \mid X=0.7) = 0 \quad \text{otherwise.}$$

Boosting

- We want to do binary classification. Cat or Dog, Red or Blue, Boy or Girl.

Boosting

- We want to do binary classification. Cat or Dog, Red or Blue, Boy or Girl.
- We have a bunch of classifiers. They all suck.

Boosting

- We want to do binary classification. Cat or Dog, Red or Blue, Boy or Girl.
- We have a bunch of classifiers. They all suck.
- More technically they all predict things correctly with probability "slightly" over a half. Weak predictors

Boosting

- We want to do binary classification. Cat or Dog, Red or Blue, Boy or Girl.
- We have a bunch of classifiers. They all suck.
- More technically they all predict things correctly with probability "slightly" over a half. Weak predictors
- We'd really like to have a single good predictor Can we combine them into a good predictor? Usually we can!

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How good a performance could we maximally get?

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How good a performance could we maximally get?
- Well, if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How good a performance could we maximally get?
- Well, if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$
- To get a correct prediction we'd need two out of three votes.

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How good a performance could we maximally get?
- Well, if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$
- To get a correct prediction we'd need two out of three votes.
- We should distribute the average of 1.8 as two votes over ninety percent of the space.

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How good a performance could we maximally get?
- Well, if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$
- To get a correct prediction we'd need two out of three votes.
- We should distribute the average of 1.8 as two votes over ninety percent of the space.
- This can in fact be constructed. So we can get 0.9 a huge boost in performance

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How bad could performance be?

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How bad could performance be?
- Again if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How bad could performance be?
- Again if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$
- To get a wrong prediction we'd need at most one out of three votes.

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How bad could performance be?
- Again if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$
- To get a wrong prediction we'd need at most one out of three votes.
- Worst case is as much as possible in the one case and the rest getting all three.

Boosting

- We have a training set and three predictors who each get things right with probability sixty-percent.
- How bad could performance be?
- Again if we set $X + i$ as the number of correct votes in sample i, then $\mathbb{E}[X_i] = 3 \times 0.6 = 1.8$
- To get a wrong prediction we'd need at most one out of three votes.
- Worst case is as much as possible in the one case and the rest getting all three.
- So $1.8 = 3 \times p + 1 \times (1p)$ which solves to $p = 0.4$. This is again easy enough to construct.

Boosting

- Having more weak predictors helps.

Boosting

- Having more weak predictors helps.
- We'd like our weak predictors to be anti-correlated, or at least not correlated.

Boosting

- Having more weak predictors helps.
- We'd like our weak predictors to be anti-correlated, or at least not correlated.
- Notion of weighting classifiers! After all they're not all going to eb the same all the time!

Boosting

- Having more weak predictors helps.
- We'd like our weak predictors to be anti-correlated, or at least not correlated.
- Notion of weighting classifiers! After all they're not all going to eb the same all the time!
- So we can think of voting as taking $sign(h_1 + h_2 + h_3 + ... + h_n)$

Boosting

- Having more weak predictors helps.
- We'd like our weak predictors to be anti-correlated, or at least not correlated.
- Notion of weighting classifiers! After all they're not all going to eb the same all the time!
- So we can think of voting as taking $sign(h_1 + h_2 + h_3 + ... + h_n)$
- Weighting looks like $sign(w_1 h_1 + w_2 h_2 + w_3 h_3 + ... + w_n h_n)$

Riddle

Let $P(x)$ be a polynomial with nonnegative integer coefficients:

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \qquad a_i \in \mathbb{Z}_{\geq 0}.$$

You must determine all of its coefficients $a_0, a_1, \ldots, a_n$. You may ask queries of the form

"What is $P(r)$?"

where r is any rational number of your choosing. After a finite number of such queries, you must reconstruct the entire polynomial. How many guesses does this take

Riddle

- 2, first guess $P(1)$ which gives you the sum of the coefficients.

Riddle

- 2, first guess $P(1)$ which gives you the sum of the coefficients.
- This is an upper bound on coefficient size (as the coefficients are positive.

Riddle

- 2, first guess $P(1)$ which gives you the sum of the coefficients.
- This is an upper bound on coefficient size (as the coefficients are positive.
- Then ask for anything larger than $P(1)$, powers of 10 are convient. For example, if $P(x) = 12x^2 + 45x1$, then $P(100) = 124501$