Домашнее задание

3.1 Lamps

Так задание на байесовский подход к оцениванию (aka "Coin") не очень зашло, но я считаю, что оно очень важное, мы его повторим (задание все равно будет засчитано единственный раз, тем, у кого я зачел предыдущую версию, переделывать не надо).

Про байесовские априорные сопряженные распределения можно прочитать на википедии, там же есть таблица: `http://en.wikipedia.org/wiki/Conjugate_prior`.

Рассмотрим сперва простой вариант задачи:

Задача. На заводе по производству лампочек наладчик настраивает станок на глаз. В результате λ — характеристика станка, влияющая на надежность выпускаемых лампочек, оказывается распределена случайно и имеет Гамма-распределение (плотность $f_\lambda(x) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta}$, Γ — гамма-функция Эйлера) с параметрами $k = 9, \theta = 1$. Считаем, что при фиксированной λ время работы лампы в минутах — случайная величина с экспоненциальным распределением с параметром λ.

Первую выпущенную лампочку проверяют на ОТК. Она сгорает ровно через минуту. Найти апостериорное распределение, величины λ, его моду, матожидание и дисперсию. *Выводить матожидание, дисперсию и моду для Гамма-распределения не надо, можно посмотреть в таблице*

Hint: для удобства лучше всего заменить в Гамма-распределении параметр θ на $\beta = 1/\theta$, как сделано на википедии.

Ответ (для самопроверки): апостериорное распределение λ — $\Gamma(k = 10, \theta = 1/2)$. Матожидание 5, дисперсия 2.5, мода 4.5.

Теперь более сложный вариант.

Задача. На заводе по производству лампочек наладчик настраивает станок на глаз. В результате λ — характеристика станка, влияющая на надежность выпускаемых лампочек, оказывается распределена случайно и имеет Гамма-распределение с параметрами $k = 9, \theta = 1$. Считаем, что при фиксированной λ время работы лампы в минутах — случайная величина с экспоненциальным распределением с параметром λ.

Затем выпускают k ламп и последовательно испытывают их. Они сгорают за $t_1, t_2, \ldots, t_k$ минут соответственно.

Промоделировать λ и $t_1, \ldots, t_k$ согласно условию задачи. Построить последовательность апостериорных байесовских распределений параметра λ после первого, второго и так далее измерений. Изобразить плотности на графике вместе с истинным значением λ. Проинтерпретировать результат.

3.2 Diabetes

В классе мы обсуждали такие понятия, как Bayesian odd и Bayesian factor. Попробуем использовать эти знания.

Задача. Пусть нам известно, что на планете Плюк 1% жителей страдают диабетом. Еще нам известно, что диабетик отличается от здорового уровнем сахара в крови: у здоровых уровень сахара имеет нормальное распределение $\mathcal{N}(\mu = 1, \sigma = 0.35)$, а у больных — $\mathcal{N}(\mu = 3, \sigma = 2)$.

Пусть мы проводим диспансеризацию. Для каждого пациента мы можем по его уровню сахара определить байесовский риск (Bayesian odd), т.е. отношение байесовской

вероятности того, что пациент болен к тому, вероятности того, что здоров и принять решение по этому значению. *Я предлагаю просто сравнивать риск с 1, на практике обычно используют более хитрую шкалу.*

Оценить (теоретически и с помощью моделирования) процент больных, признанных здоровыми, и процент здоровых, признанных больными (эти меры называются false negative rate и false positive rate, подробнее смотреть, например, здесь: `http://en.wikipedia.org/wiki/Precision_and_recall`). Найти decision bound, т.е. такое пограничное значение уровня сахара, при превышении которого мы считаем пациента диабетиком.

Я понимаю, что "нормально распределенный уровень сахара" это полный треш и все такое. Вполне можно использовать более сложные распределения (например, Гамма), главное, чтобы плотности не очень сильно перекрывались. Если кто-то найдет реальные сведения об уровне сахара у диабетиков и подберет распределения так, что они будут примерно похожи на реальность, он будет большой молодец. Имейте ввиду, что если взять слишком сложные распределения, то decision bound может состоять более чем из одной точки.