1) Указания на сентябрь, полученные от научного руководителя

Нужно было изучить интересные для себя книги и статьи (например, используя Google Scholar) и выбрать ту книгу или статью, на которую планируется опираться в процессе самостоятельной работы и создания курсовой. Также нужно было выбрать тему курсовой и обозначить план работ в рамках выбранной темы.

2) Мой План работы на сентябрь 2020

Подумать о том, какие из направлений, уже освещенных в процессе обучения теории вероятностей, математической статистике, статистическому практикуму, актуарной и финансовой математике мне наиболее понравились? Какие из них показались непонятными, но интересными? Может быть, с чем-то непонятным захотелось разобраться и реализовать это на практике? Как только такая тема находится, то найти по этой теме книги и статьи, а потом выбрать из них подходящую.

3) Что конкретно сделано за сентябрь 2020 из намеченного

Я подумала и поняла, что самым непонятным был уход на статистическом практикуме от метода максимального правдоподобия (ММП) (замечательные свойства которого долго и упорно изучались на математической статистике) к методу максимума апостериорной вероятности (MAP), а потом к моделям со скрытыми переменными. Мне захотелось получить математическое обоснование того, почему вообще такие алгоритмы хорошо работают, и в каких случаях и задачах их нужно выбирать.

Были изучены 2 статьи и рассмотрены 2 книги этих же авторов:

Kevin P. Murphy “Machine Learning: A Probabilistic Perspective” ISBN 978-0-262-01802-9

Christopher M. Bishop “Pattern recognition and Machine Learning” ISBN-10: 0-387-31073-8 ISBN-13: 978-0387-31073-2

В качестве основной литературы выбрана первая книга.

Начато изучение главы 3 из этой книги.

Выбрана следующая тема курсовой: “О применимости непараметрической модели, параметрической модели и модели со скрытыми переменными в задаче о восстановлении плотности”

Направление работы: разобраться, почему и как 3 вышеописанных модели работают, доказать это все математически и продемонстрировать эту работу на примерах.

4) Что не сделано за сентябрь 2020 из намеченного

Все намеченное на сентябрь 2020 сделано.

5) Причины

Все сделано.

6) Что сделано из того, что не было запланировано

Составлен детальный план курсовой.

Название: “О применимости непараметрической модели, параметрической модели и модели со скрытыми переменными в задаче о восстановлении плотности”

План:
Пусть нам дан какой-то вектор наблюдений $X = (x_1, x_2, \dots, x_N)$, и мы хотим понять, из какого распределения пришли эти наблюдения, то есть нужно восстановить плотность $p(x)$. Зачем нам может понадобиться восстановить плотность – это отдельный вопрос. Например, для построения оптимального байесовского классификатора, который работает по формуле
$a^*(x) = \underset{y \in Y}{\operatorname{argmax}} p(y|x) = \underset{y \in Y}{\operatorname{argmax}} \frac{p(x|y)p(y)}{p(x)} = \underset{y \in Y}{\operatorname{argmax}} p(x|y)p(y).$
Вот тут-то нам и надо при каждом значении класса $y = y_0$ восстановить плотность $p(x|y_0)$. Но об этом классификаторе – как-нибудь в другой раз, а пока мы хотим научиться восстанавливать плотность по наблюдениям.

1) Самый простой способ (это и есть непараметрическая модель) получить представление о $p(x)$ – это построить гистограмму. Но проблема в том, что это будут столбики, то есть не будет дифференцируемости плотности, а нам бы этого очень хотелось. Эта проблема решается методом ядерного сглаживания. Приводим пример сглаживания с ядром в виде нормального распределения. Вставить картинку с гистограммой-столбиками и результатом после сглаживания.

2) Второй по сложности способ – это предположить, что распределение лежит в каком-то фиксированном классе распределений (этот класс мы сами назначаем исходя из каких-то своих представлений о происхождении данных), и попытаться оценить параметр θ нашего распределения.
Делать это можно по-разному.

■ Во-первых, можно применить метод максимума правдоподобия (далее ММП):
$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} p(x|\theta) = \underset{\theta}{\operatorname{argmax}}(\log p(x|\theta))$
Для этого надо приравнять к нулю производную от правдоподобия $L(\theta) = p(X|\theta) = \prod_{i=1}^{N} p(X_i|\theta)$. Но ведь таким образом максимум мы найдем только у вогнутой функции. Поэтому доказываем, что если распределение принадлежит экспоненциальному семейству $p(x|\theta) = \frac{f(x)}{g(\theta)} e^{\theta^T u(x)}$, где $g(\theta) = \int f(x)\, e^{\theta^T u(x)} dx$, то $\log L(\theta)$ будет функцией вогнутой. Ну как это сделать: $\log L(\theta) = \log p(x|\theta) = \log \frac{f(x)}{g(\theta)} e^{\theta^T u(x)} = \sum_{i=1}^{N} f(x_i) - N\, log\, g(\theta) + \theta^T \sum_{i=1}^{N} u(x_i)$. И доказать, что вторая производная по θ всегда отрицательно определена. Окажется, что $\frac{\partial^2 g(\theta)}{\partial \theta_i \partial \theta_j} = cov(u_i, u_j)$, а поскольку матрица ковариаций всегда неотрицательно определена, то мы докажем вогнутость. Явно посчитать ММП для монетки и для нормального распределения. Явно доказать, что все знакомые распределения (в частности, нормальное, экспоненциальное, пуассоновское, равномерное, бета-распределения, биномиальное, категориальное, распределения Дирихле) лежат в экспоненциальном семействе.

Остается вопрос: что делать в случае не экспоненциального семейства?
И что делать в случае маленьких N? Ведь хорошие свойства ММП (а именно несмещенность, состоятельность и неотрицательная определенность) есть только для больших N, а для маленьких нет. Например, если $N = 3$, и мы при 3 бросках монетки с вероятностью θ выпадения орла все 3 раза получили решку, то неужели мы верим, что выдаваемая методом ММП оценка $\theta^*_{ML} = \frac{\sum_{i=1}^{N} x_i}{N} = 0$ очень хороша, то есть правда ли монетка совсем не может выпадать орлом? Ну конечно нет! Эту проблему можно попытаться исправить регуляризацией (то есть ввести $a + b$ фиктивных бросков, из которых a раз выпал орел, и получить $\theta^*_{ML,reg} = \frac{\sum_{i=1}^{N} x_i + a}{N+a+b}$). Но непонятно, ни то, откуда мы взяли такой способ регуляризации, ни то, как действовать в общем случае; А что делать в случае не экспоненциального семейства – вообще не понятно (ответ: применять модель со скрытыми переменными). Оказывается, что ответ на первый вопрос (то есть если семейство экспоненциальное, но по ML получается не очень хорошая оценка) дает метод апостериорной вероятности (далее MAP).

■ Как работает MAP:
По формуле Байеса: $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$.
Мы сами (исходя из вида $p(x|\theta)$) выбираем $p(\theta)$ таким, как нам удобно. Нам было бы очень удобно, если бы $p(x|\theta)p(\theta)$ лежало в таком же классе, что и $p(\theta)$. Такие распределения называются сопряженными. Поэтому имея $p(x|\theta)$, мы выберем в качестве $p(\theta)$ сопряженное к $p(x|\theta)$ распределение.
Дать обзор сопряженных распределений. Явно доказать, что сопряженным к биномиальному распределению является бета-распределение, а сопряженным к категориальному является распределение Дирихле. Посчитать явно θ_{MAP} для монетки и сравнить с θ_{ML} для монетки. Убедиться, что $\theta_{MAP} = \theta^*_{ML,reg}$.

3) Что делать, если $p(x|\theta)$ не из экспоненциального семейства? Если по данным видно, что они не из экспоненциального семейства, то есть что они похожи на смесь распределений из экспоненциального семейства (например, смесь гауссиан), то есть одним классом распределений (пусть даже мы и найдем наиболее подходящий параметр θ для выбранного класса с помощью ML или MAP) мы его не опишем, то тогда надо вводить скрытые переменные (в случае смеси гауссиан скрытые переменные z_1 и z_2 отвечают за то, к какой гауссиане относятся наблюдения) и применять EM-алгоритм.
Подчеркнуть, что скрытые переменные мы выбираем сами, поэтому мы их так можем выбрать, чтобы совместное распределение $p(X, Z|\,\theta)$ было из экспоненциального семейства, тогда $\log p(X, Z|\,\theta)$ будет вогнутой, а значит, и $E_z \log p(X, Z|\,\theta)$ будет вогнутой, и мы сможем найти ее максимум с помощью ML или MAP. Замечание: интеграл (то есть математическое ожидание) даже можно не считать аналитически, а применять методы Монте-Карло.
Описать EM-алгоритм, для этого ввести понятия энтропии, вариационной нижней оценки, дивергеницию Кульбака-Лейблера, убедиться, что она не метрика, но неотрицательна. Используя эту неорицательность, найти вариационную нижнюю оценку и доказать EM-алгоритм. Отметить, что шаг E этого EM-алгоритма является мягкой кластеризацией, и в пределе превращается в жесткую кластеризацию, то есть в KNN. Нарисовать картинку-иллюстрацию к EM-алгоритму. Использовать этот алгоритм для разделения гауссиан. Использовать его для примера с двумя монетками. Использовать его для примера узнавания правильного ответа исходя из ответов студентов.

Мораль: плотность мы научились восстанавливать.

Далее можно двигаться в нескольких направлениях:
1) Применить модель со скрытыми переменными для LDA (Latent Dirichlet Allocation) и реализовать ее (написать программу и убедиться, что она работает)
2) Применить модель со скрытыми переменными для построения вариационного автоэнкодера.
3) Применить какие-то (или все) модели восстановления плотности для построения оптимального байесовского классификатора. В частности, напрограммировать модель наивного байесовского классификатора, который предполагает, что компоненты независимы при условии таргета. Дальше сказать, что это предположение наивности не всегда работает (например, зарплата и место работы явно коррелированы между собой). Чтобы определить, применимо ли это наивное предположение, используются статистические тесты. И пойти дальше исследовать методы проверки статистических гипотез.
4) После построения оптимального байесовского классификатора пойти делать сравнительный анализ классификаторов (то есть генеративные, дискриминативные вероятностные и просто дискриминативные модели)
5) Пойти рассматривать методы Монте-Карло.

7) План работы на октябрь 2020

Получить одобрение научного руководителя по теме и плану курсовой.
Дочитать главу 3 из книги.
Разобраться, что такое сопряженные распределения и записать/напечатать явную проверку сопряженности биномиального-бета и категориального-Дирихле.
Явно записать процедуры ML и MAP для монетки и сравнить ответы, сделать замечание про регуляризацию.
Все результаты набрать в Microsoft Word.