

Министерство науки и высшего образования Российской
Федерации Федеральное государственное бюджетное
образовательное учреждение высшего образования

Отчет по дисциплине

«Физика» по теме:

**«Предсказание оптических свойств материалов на
основе регрессионных моделей»**

Выполнил студент группы

ИСИ6-23-1

Шифр группы

Подпись

Р.А. Затопляев

И.О. Фамилия

Проверил преподаватель

Подпись
Фамилия

Д.С. Зограбян
И.О.

Иркутск 2025 г.

Предсказание оптических свойств материалов на основе регрессионных моделей

Аннотация

В данной работе исследуется возможность предсказания коэффициента преломления (n) простых неорганических материалов на основе их физических параметров: плотности, ширины запрещённой зоны и диэлектрической проницаемости. Используются данные из базы [Materials Project](#) и один из методов машинного обучения (модель случайного леса).

Введение

Актуальность: коэффициент преломления — ключевая характеристика в оптике, фотонике, лазерных технологиях. Экспериментальное измерение n трудоёмко, поэтому важно разработать вычислительные метод.

Цель: разработать модель случайного леса, которая будет предсказывать коэффициент преломления на основе физических свойств материала.

Задачи:

- 1) Собрать данные о простых неорганических материалах с сайта [Materials Project](#).
- 2) Собрать данные о настоящих коэффициентах преломления, полученных экспериментально, с базы данных [Refractive.index-database](#).
- 3) Сгенерировать синтетические данные для улучшения качества модели.
- 4) Обработать данные. Подготовить их для анализа.
- 5) Обучить модель, сделать анализ полученных результатов.

ГЛАВА 1. Теоретическая часть

1.1 Физические основы

Дадим определение для всех анализируемых параметров, используемых в модели, и опишем их влияние на коэффициент преломления материала.

Определение: Коэффициент (показатель) преломления среды n - безразмерная величина, которая показывает, во сколько раз скорость v света в веществе меньше, чем скорость c в вакууме.

$$n = \frac{c}{v}$$

Определение: Запрещённая зона - тот диапазон энергий, в котором электрон не может находиться, потому что такие состояния для него запрещены.

Чем легче электроны возбуждаются (то есть чем меньше ширина запрещенной зоны), тем сильнее отклик материала на свет, и тем выше коэффициент преломления.

Определение: Диэлектрическая проницаемость среды — это физическая величина, которая описывает способность материала накапливать электрический заряд в ответ на приложенное электрическое поле. В простых словах, диэлектрическая проницаемость характеризует, насколько материал может «поглощать» или «ослаблять» электрическое поле, проходящее через него.

Существуют также ионная, электронная и еще несколько видов диэлектрических проницательностей, которые вносят вклад в полную диэлектрическую проницательность.

Материалы с высокой диэлектрической проницаемостью сильнее взаимодействуют с электрическим компонентом электромагнитных волн, замедляя их и изменяя их траекторию.

Определение: плотность материала – это физическая величина, которая характеризует массу вещества в единице объема.

В материалах с высокой плотностью молекулы или атомы располагаются теснее, что приводит к более высокому уровню взаимодействия с электрическим полем света. Это снижает скорость распространения света в материале, так как диэлектрическая проницаемость увеличивается, что замедляет свет.

1.2 Машинное обучение в физике

Процесс построения модели случайного леса:

- 1) Подготовка данных. Для упрощения создания модели случайного леса, будем использовать данные о полупроводниковых материалах. Будем использовать данные с сайта [Materials Project](#) и с сайта [Refractive index](#). Соберем необходимые данные в таблицу, а именно: плотность, ширина запрещенной зоны и электронная диэлектрическая проницаемость. Электронная наиболее важна для описания свойств полупроводников, а также соберем значения показателей коэффициентов преломления, полученных экспериментально, чтобы проверить корректность нашей модели.
- 2) Предобработка данных. Диэлектрическая проницаемость представлена в базе данных как тензор 3 на 3. Тензор представляет собой матрицу 3 на 3, которая описывает как материал реагирует на электрические поля в разных направлениях. Нужно добавить в анализируемые параметры электронную диэлектрическую проницаемость для каждого направления (взять диагональные элементы). Чтобы улучшить качество модели случайного леса, прологарифмируем параметры – это нужно для уменьшения “колебания” данных.

3) Построение модели случайного леса.

Случайный лес – это алгоритм, который строит множество деревьев решений и объединяет их результаты.

Для каждого дерева в лесу:

- 1) Выбираем случайно подмножество признаков из $\log(\rho), \log(E + 10^{-6}), \log(\epsilon_x), \log(\epsilon_y), \log(\epsilon_z)$. Размером квадратный корень из количества признаков.
 ρ - плотность материала
 E – ширина запрещенной зоны
 $\epsilon_x, \epsilon_y, \epsilon_z$ – компоненты тензора электронной диэлектрической проницаемости
- 2) Создаем бутстрап выборку из обучающих данных.
- 3) Процесс построения узла дерева происходит на основе определения условия разделения. Условие разделения определяется с помощью минимизации суммы квадратов ошибок (RSS). То есть ищем условие, при котором узлы дерева построятся с наименьшей суммой квадратов ошибок.
- 4) Дерево строится по этой выборке до максимальной глубины или пока не будет выполнено условие остановки.
- 5) Для случайного леса строим множество таких деревьев (в нашем случае 200).
- 6) После чего происходит агрегация результатов, берем среднее из всех предсказаний леса.

1.3 Программная реализация случайного леса для предсказания коэффициента преломления полупроводников.

Таблица 1

Пример данных, подготовленных для обучения:

material_id	formula	refractive_index	log_density	log_band_gap	log_eps_xx	log_eps_yy	log_eps_zz
mp-632291	H2	1.0001403992	-1.9513333480475656	2.180406272516592	0.3828386222738061	0.3828386222738061	0.5070509245443062
mp-850274	H2	1.0001403992	-2.3344101594595927	2.1440348660031847	0.5300941084333929	0.3913980899133664	0.3913980899133664
mp-542605	Se	1.4673	1.3794236356562646	0.3294483410049499	1.7065070122920645	1.7375464569604178	1.3522316188929857
mp-14	Se	1.4673	1.5030970593546849	-0.1318174337201865	2.081014166010102	2.0810139782009	2.585075140682012
mp-19	Te	2.937	1.7707399821252985	-1.6841560707135173	4.466299853407127	4.466300440426263	5.199452690555336
mp-567313	Te	2.937	1.766168369419305	-0.578745856449972	3.746361839660141	3.746361645452648	4.09674913525112
mp-990424	C	2.274911838790932	-0.4577991319260737	-4.389978164461582	1.3881592735891313	2.679678883641845	0.1869059940727534
mp-611219	Hg	1.3008964469048818	1.3559931379519805	0.565939181645667	0.363895320171204	1.434184267057506	1.4911353923638713
mp-12103	N2	1.000296298788009	-0.471554078246063	-0.0371817532455363	0.3420772129387127	0.3420770509631115	0.3423179101869637
mp-165	Si	4.299202898550725	0.8403986883613801	-0.8234814038672156	2.4903630183642123	2.522728493954485	2.618536031465605
mp-23907	H2	1.0001403992	-3.219468752776016	2.0128744310118183	0.2320879726284641	0.2320881189524489	0.2321677029006938
mp-570752	H2	1.0001403992	-3.281751484356535	1.892494453591236	0.2194955372174285	0.2194951682282978	0.2247357071611461
mp-154	N2	1.000296298788009	-0.0552552173169806	1.9934752093498584	0.4250103754927365	0.4250103754927365	0.4250103754927365
mp-672234	N2	1.000296298788009	-0.6382235232303359	1.095674978994699	0.3334145165563443	0.3334145165563443	0.3334145165563443
mp-999498	N2	1.000296298788009	1.2367490720741785	1.3760927511987815	1.5409526713667376	1.5409526713667376	1.5409526713667376
mp-864900	Hg	1.3008964469048818	2.4624946292124816	-0.6810190248699667	2.9325046741163057	2.9325046741163057	2.9325046741163057
mp-971661	Si	4.299202898550725	0.7481432932352006	0.316416020031128	2.488441849432949	2.488441849432949	2.488441849432949

refractive_index – настоящий коэффициент преломления

log_density – логарифм плотности материала

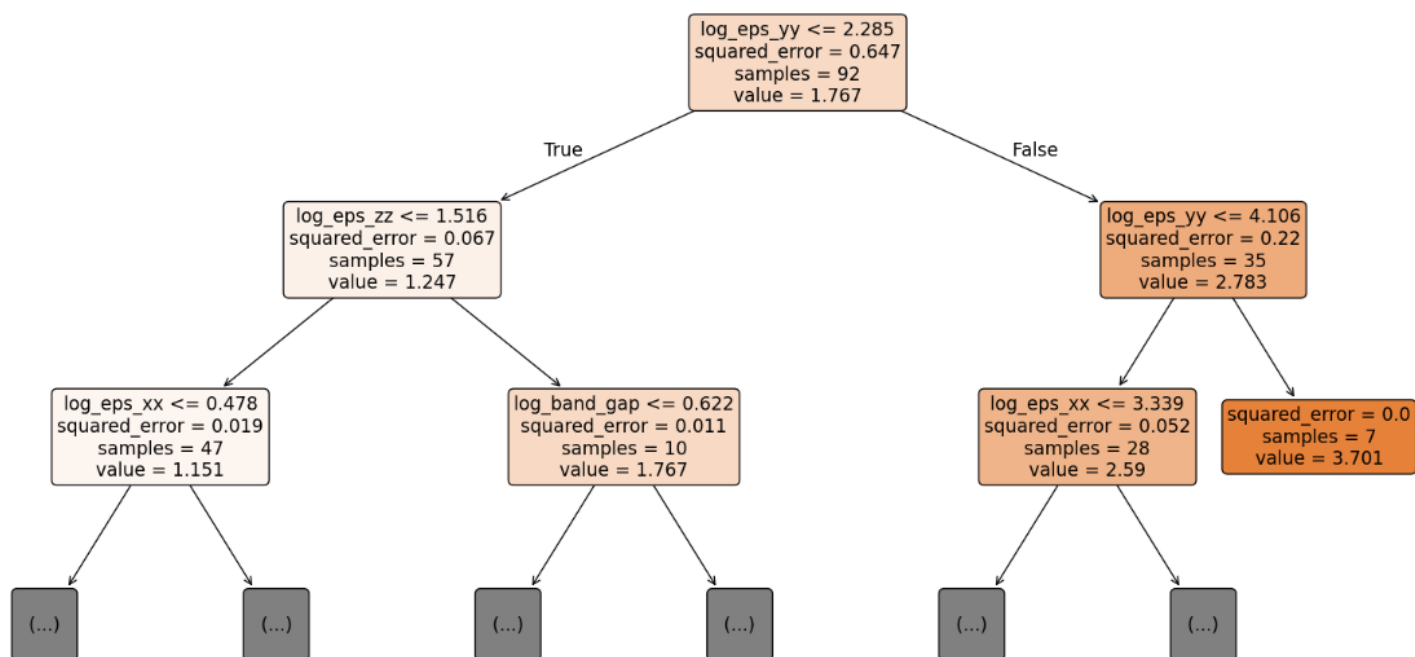
log_band_gap – логарифм ширины запрещенной зоны материала

log_eps_xx, log_eps_yy, log_eps_zz – логарифмы компонент электронной диэлектрической проницаемости материала

```
# Загрузка данных
files = ['materials_refractive_index.csv', "synthetic_realistic_data.csv"]
df_list = [pd.read_csv(f) for f in files]
df_train = pd.concat(df_list, ignore_index=True)
# Признаки для обучения
features = ['log_density', 'log_band_gap', 'log_eps_xx', 'log_eps_yy', 'log_eps_zz']
# Необходимо определить коэффициент преломления
target = 'refractive_index'

X = df_train[features]
y = df_train[target]
# Разделение на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Обучение модели
model = RandomForestRegressor(n_estimators=200, max_depth=15, min_samples_leaf=5, random_state=42)
model.fit(X_train, y_train)
```

Пример построенного дерева из случайного леса:




```
# Протестируем модель на тестовой выборке
y_pred = model.predict(X_test)
# Сравнение предсказанных и реальных значений
comparison = pd.DataFrame({
    'true_n': y_test.values,
    'predicted_n': y_pred
})
comparison['absolute_error'] = abs(comparison['true_n'] -
comparison['predicted_n'])
```

Таблица 2

Результаты тестирования:

true_n	predicted_n	absolute_error
2.4755835019686687	2.504592156452922	0.02900865448425316
1.1916067483449126	1.2455712399664092	0.053964491621496524
2.187667773807086	2.13528115288936	0.05238662091772639
2.191703646638726	2.223634338296681	0.031930691657954924
2.1584632137791067	2.1991969994890015	0.04073378570989483
2.024210709947327	1.9498959044860704	0.0743148054612568
2.754653786684641	2.8508168181946782	0.09616303151003747
1.6064901329008	1.5765266148305423	0.029963518070257544
3.266767525130231	3.3139390955953956	0.047171570465164514
2.991154047666298	2.902048690299554	0.08910535736674374
2.6168244089300554	2.6709701174394076	0.054145708509352186
1.6435853210869311	1.6672562679120488	0.023670946825117722
2.1147424877928813	2.3313567064741747	0.21661421868129338
2.2566138724947806	2.311566777092631	0.054952904597850516
2.459553557285805	2.556676701684002	0.0971231443981968
1.849082642378884	1.878610842978645	0.029528200599761023
2.975181144739365	3.033167958661542	0.05798681392217686

Интерпретация результатов:

Случайный лес показал хорошую способность прогнозировать коэффициент преломления материалов на основе физических характеристик.

Модель может быть полезна для ускоренного анализа материалов с заданными оптическими свойствами без необходимости проведения затратных экспериментов.

Также до модели случайного леса была попытка построения линейной регрессионной модели. Была видна явная зависимость коэффициента преломления от анализируемых оптических свойств.

- Плотность. Чем выше плотность тем меньше n
- Электронная диэлектрическая проницаемость. Чем она выше, тем больше n .
- Ширина запрещенной зоны. Чем она меньше, тем больше n

Но после анализа полученной модели, стало ясно, что она не пригодна для нашей задачи, так как коэффициент преломления не зависит линейно от анализируемых свойств материалов.

ВЫВОДЫ

Проделав, данную проектную работу, я значительно улучшил свои знания в области физических структур материалов, на практике узнал насколько сильно влияют различные параметры материала на его коэффициент преломления.

Также после выполнения данного исследования были получены новые навыки в области машинного обучения, а именно построения линейной регрессионной модели и модели случайного леса.

Полученная в результате работы модель далеко не идеальна и требует различного рода доработок, но я думаю, что она имеет право на дальнейшее существование.