

Классификация изображений (распознавание рукописных цифр)

Семинар в БГУ
11 сентября 2012 года

Содержание презентации

- ▶ Почему распознавание рукописных цифр
- ▶ Описание данных
- ▶ Предварительный анализ данных. Визуализация
- ▶ Уменьшение размерности данных для визуализации
- ▶ Классификация
- ▶ Уменьшение размерности данных для классификации
- ▶ Демонстрационное видео



Почему только цифры

[illegible]

Мр Вил, ген В, С Олиф
Крово Ес 4,5-10¹²кв 140,47
L 6,1 x 10¹² рд 75
моча рд все 1018
с-бубр
клет ери оа
с-бубр гл имбромин
гелс
с-бубр олимф.
с-бубр олимф
с-бубр, 0,5% рд
с-бубр.
с-бубр 45
с-бубр

[illegible]

в ксер/ксер.
231.

310 x 111. met.

 34400.

113
 114
 115
 116
 117
 118
 119
 120
 121
 122
 123
 124
 125
 126
 127
 128
 129
 130
 131
 132
 133
 134
 135
 136
 137
 138
 139
 140
 141
 142
 143
 144
 145
 146
 147
 148
 149
 150
 151
 152
 153
 154
 155
 156
 157
 158
 159
 160
 161
 162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173
 174
 175
 176
 177
 178
 179
 180
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191
 192
 193
 194
 195
 196
 197
 198
 199
 200
 201
 202
 203
 204
 205
 206
 207
 208
 209
 210
 211
 212
 213
 214
 215
 216
 217
 218
 219
 220
 221
 222
 223
 224
 225
 226
 227
 228
 229
 230
 231
 232
 233
 234
 235
 236
 237
 238
 239
 240
 241
 242
 243
 244
 245
 246
 247
 248
 249
 250
 251
 252
 253
 254
 255
 256
 257
 258
 259
 260
 261
 262
 263
 264
 265
 266
 267
 268
 269
 270
 271
 272
 273
 274
 275
 276
 277
 278
 279
 280
 281
 282
 283
 284
 285
 286
 287
 288
 289
 290
 291
 292
 293
 294
 295
 296
 297
 298
 299
 300
 301
 302
 303
 304
 305
 306
 307
 308
 309
 310
 311
 312
 313
 314
 315
 316
 317
 318
 319
 320
 321
 322
 323
 324
 325
 326
 327
 328
 329
 330
 331
 332
 333
 334
 335
 336
 337
 338
 339
 340
 341
 342
 343
 344
 345
 346
 347
 348
 349
 350
 351
 352
 353
 354
 355
 356
 357
 358
 359
 360
 361
 362
 363
 364
 365
 366
 367
 368
 369
 370
 371
 372
 373
 374
 375
 376
 377
 378
 379
 380
 381
 382
 383
 384
 385
 386
 387
 388
 389
 390
 391
 392
 393
 394
 395
 396
 397
 398
 399
 400
 401
 402
 403
 404
 405
 406
 407
 408
 409
 410
 411
 412
 413
 414
 415
 416
 417
 418
 419
 420
 421
 422
 423
 424
 425
 426
 427
 428
 429
 430
 431
 432
 433
 434
 435
 436
 437
 438
 439
 440
 441
 442
 443
 444
 445
 446
 447
 448
 449
 450
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460
 461
 462
 463
 464
 465
 466
 467
 468
 469
 470
 471
 472
 473
 474
 475
 476
 477
 478
 479
 480
 481
 482
 483
 484
 485
 486
 487
 488
 489
 490
 491
 492
 493
 494
 495
 496
 497
 498
 499
 500
 501
 502
 503
 504
 505
 506
 507
 508
 509
 510
 511
 512
 513
 514
 515
 516
 517
 518
 519
 520
 521
 522
 523
 524
 525
 526
 527
 528
 529
 530
 531
 532
 533
 534
 535
 536
 537
 538
 539
 540
 541
 542
 543
 544
 545
 546
 547
 548
 549
 550
 551
 552
 553
 554
 555
 556
 557
 558
 559
 560
 561
 562
 563
 564
 565
 566
 567
 568
 569
 570
 571
 572
 573
 574
 575
 576
 577
 578
 579
 580
 581
 582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593
 594
 595
 596
 597
 598
 599
 600
 601
 602
 603
 604
 605
 606
 607
 608
 609
 610
 611
 612
 613
 614
 615
 616
 617
 618
 619
 620
 621
 622
 623
 624

Почему только цифры

Дата посещения	Амбулатор. на дому (вписать)	Жалобы больного, объективные данные, течение и диагноз болезни, подписи врачей и консультантов	Назначения и отметки о выдаче листа нетрудоспособности
11/05		Грипп	
		Состояние	
		температура	
		пульс	
		давление	
		сердце	
		легкие	
		желудок	
		кишечник	
		мочеполовая	
		нервная	
		общее состояние	
		прогноз	

изобра
Мр Вил, ген В, С Ош
Крово Ес 4,5-10 кмк 142,77
L 6,1 x 10⁵ POF 5
моча ур все 1018
С-бубз
лент ери ов
С-тв гу мндромин
meus
Оликоиди,
Фрикеити
таба, 0,5% РР
слудин.
м тирами и S
мксатрас
113

[illegible]

в ксер/нечер.
231.

310 x 111. met.

 34410.

112
 113
 114
 115
 116
 117
 118
 119
 120
 121
 122
 123
 124
 125
 126
 127
 128
 129
 130
 131
 132
 133
 134
 135
 136
 137
 138
 139
 140
 141
 142
 143
 144
 145
 146
 147
 148
 149
 150
 151
 152
 153
 154
 155
 156
 157
 158
 159
 160
 161
 162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173
 174
 175
 176
 177
 178
 179
 180
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191
 192
 193
 194
 195
 196
 197
 198
 199
 200
 201
 202
 203
 204
 205
 206
 207
 208
 209
 210
 211
 212
 213
 214
 215
 216
 217
 218
 219
 220
 221
 222
 223
 224
 225
 226
 227
 228
 229
 230
 231
 232
 233
 234
 235
 236
 237
 238
 239
 240
 241
 242
 243
 244
 245
 246
 247
 248
 249
 250
 251
 252
 253
 254
 255
 256
 257
 258
 259
 260
 261
 262
 263
 264
 265
 266
 267
 268
 269
 270
 271
 272
 273
 274
 275
 276
 277
 278
 279
 280
 281
 282
 283
 284
 285
 286
 287
 288
 289
 290
 291
 292
 293
 294
 295
 296
 297
 298
 299
 300
 301
 302
 303
 304
 305
 306
 307
 308
 309
 310
 311
 312
 313
 314
 315
 316
 317
 318
 319
 320
 321
 322
 323
 324
 325
 326
 327
 328
 329
 330
 331
 332
 333
 334
 335
 336
 337
 338
 339
 340
 341
 342
 343
 344
 345
 346
 347
 348
 349
 350
 351
 352
 353
 354
 355
 356
 357
 358
 359
 360
 361
 362
 363
 364
 365
 366
 367
 368
 369
 370
 371
 372
 373
 374
 375
 376
 377
 378
 379
 380
 381
 382
 383
 384
 385
 386
 387
 388
 389
 390
 391
 392
 393
 394
 395
 396
 397
 398
 399
 400
 401
 402
 403
 404
 405
 406
 407
 408
 409
 410
 411
 412
 413
 414
 415
 416
 417
 418
 419
 420
 421
 422
 423
 424
 425
 426
 427
 428
 429
 430
 431
 432
 433
 434
 435
 436
 437
 438
 439
 440
 441
 442
 443
 444
 445
 446
 447
 448
 449
 450
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460
 461
 462
 463
 464
 465
 466
 467
 468
 469
 470
 471
 472
 473
 474
 475
 476
 477
 478
 479
 480
 481
 482
 483
 484
 485
 486
 487
 488
 489
 490
 491
 492
 493
 494
 495
 496
 497
 498
 499
 500
 501
 502
 503
 504
 505
 506
 507
 508
 509
 510
 511
 512
 513
 514
 515
 516
 517
 518
 519
 520
 521
 522
 523
 524
 525
 526
 527
 528
 529
 530
 531
 532
 533
 534
 535
 536
 537
 538
 539
 540
 541
 542
 543
 544
 545
 546
 547
 548
 549
 550
 551
 552
 553
 554
 555
 556
 557
 558
 559
 560
 561
 562
 563
 564
 565
 566
 567
 568
 569
 570
 571
 572
 573
 574
 575
 576
 577
 578
 579
 580
 581
 582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593
 594
 595
 596
 597
 598
 599
 600
 601
 602
 603
 604
 605
 606
 607
 608
 609
 610
 611
 612
 613
 614
 615
 616
 617
 618
 619
 620
 621
 622
 623

Почему только цифры

Дата посещения	Амбулатор на дому (вписать)	Жалобы больного, объективные данные, течение и диагностика болезни, подписи врачей и консультантов	Назначения и отметки о выдаче листа нетрудоспособности
10.05		Состояние улучшилось мимическая мускулатура мышцы шеи гортани языка глотки пищеварительного тракта дыхательного аппарата	

Клеи

Соиссиси мими

МИМИМИМИ МИ МИМИ

МИМИМИМИ

МИМИМИМИ

ЦМММММ

МИМИМИМИ

No1

Почему только цифры

Дата посещения	Амбулатор. на дому (вписать)	Жалобы больного, объективные данные, течение и диагноз болезни, подписи врачей и консультантов	Назначения и отметки о выдаче листа нетрудоспособности
10/05	Сосисим или	пилим месе	семишнее
	тип же	слишишь	мимими
	Фглиссоесеев	мля	

Сосисим или
пилим месе
семишнее
тип же
слишишь
мимими

Фглиссоесеев
мля

№2

Почему только цифры

Дата посещения	Амбулатор. на дому (вписать)	Жалобы больного, объективные данные, течение и диагноз болезни, подписи врачей и консультантов	Назначения и отметки о выдаче листка нетрудоспособности
10/05	Григорьев	Состояние удовлетворительное	
		Минимизируемые	
		Минимизируемые	
		Минимизируемые	
		Минимизируемые	
		Минимизируемые	
		Минимизируемые	
		Минимизируемые	
		Минимизируемые	
		Минимизируемые	

Били
Соитии инь
Пили и мясо
шиш шее
тип янь
титя
слиплись
чтим мясо
ишемия
зриши ось
я илия

№3

Почему только цифры

Дата посещения	Амбулатор. на дому (вписать)	Жалобы больного, объективные данные, течение и диагноз болезни, подписи врачей и консультантов	Назначения и отметки о выдаче листка нетрудоспособности
10.05	Жили	Солнечный инь.	
	Жили ж лень.		
	Шишки		
	Жми янь.		
	Тили.		
	Сплин мы.		
	Мишки.		
	з. Ргилавеев		
	я — лис		

Клею.
Солнечный инь.
Жили ж лень.
Шишки
Жми янь.
Тили.
Сплин мы.
Мишки.

з. Ргилавеев
я — лис

Почему только цифры

[illegible]

изорж
МВ Вил, ген В, С смр
Криво Е 4,5-10 мм 14242
L 6,1 x 10³ P 075
мока др вое 1018
с-бубр
кент ерм оз
с-бубр гут мидромм
mees
олик олимп,
одрок олимп
таб, 0,5%. РР
супрм.
м трамис и 5
мк олимп

[illegible]

матр. ест.
в кспр/нечер.

231.

hallot

310 км. мет.

mosses etc.
577 p.

112
 113
 114
 115
 116
 117
 118
 119
 120
 121
 122
 123
 124
 125
 126
 127
 128
 129
 130
 131
 132
 133
 134
 135
 136
 137
 138
 139
 140
 141
 142
 143
 144
 145
 146
 147
 148
 149
 150
 151
 152
 153
 154
 155
 156
 157
 158
 159
 160
 161
 162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173
 174
 175
 176
 177
 178
 179
 180
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191
 192
 193
 194
 195
 196
 197
 198
 199
 200

Описание данных

Данные подготовлены

- ▶ Цифры локализованы, центрированы, нормированный размер
 - ▶ готовые алгоритмы из популярных библиотек обработки изображений (OpenCV, ImageJ, AForge.NET)
- ▶ Размер одной цифры 28×28 пикселей, 256 оттенков серого
- ▶ Весь набор данных в текстовом файле, формат csv
- ▶ Первый столбец — обозначения классов



[illegible]

Предварительный анализ данных.

Визуализация

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	41	149	156	179	254	254	201	119	46	0	0	0
0	0	0	0	13	147	241	253	253	254	253	253	253	253	245	160	0	0
0	0	0	31	224	253	253	180	174	175	174	174	174	174	223	247	145	6
0	0	7	197	254	253	165	2	0	0	0	0	0	0	12	102	184	16
0	0	152	253	254	162	18	0	0	0	0	0	0	0	0	0	0	0
0	0	235	254	158	15	0	0	0	0	0	0	0	0	0	0	0	0
0	74	250	253	15	0	0	0	16	20	19	0	0	0	0	0	0	0
7	199	253	253	0	0	25	130	235	254	247	145	6	0	0	0	0	0
20	253	253	177	100	219	240	253	253	254	253	253	125	0	0	0	0	0
5	193	253	253	254	253	253	200	155	155	238	253	229	23	0	0	0	0
0	61	249	254	241	150	30	0	0	0	215	254	254	58	0	0	0	0
0	0	36	39	30	0	0	0	0	0	214	253	234	31	0	0	0	0
0	0	0	0	0	0	0	0	0	41	241	253	183	0	0	0	0	0
0	0	0	0	0	0	0	0	0	201	253	253	102	0	0	0	0	0
0	0	0	0	0	0	0	0	114	254	253	154	5	0	0	0	0	0
0	0	0	0	0	0	0	62	254	255	241	30	0	0	0	0	0	0
0	0	0	0	0	10	118	235	253	249	103	0	0	0	0	0	0	0
0	55	81	0	102	211	253	253	253	135	0	0	0	0	0	0	0	0
0	79	243	234	254	253	253	216	117	0	0	0	0	0	0	0	0	0
0	48	245	253	254	207	126	27	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



Визуализация

$\infty \rightarrow 6950210$
 $\infty \rightarrow 5040210$
 $\infty \rightarrow 4030210$
 $\infty \rightarrow 3020210$
 $\infty \rightarrow 2010210$
 $\infty \rightarrow 1000210$
 $\infty \rightarrow 0990210$
 $\infty \rightarrow 0880210$
 $\infty \rightarrow 0770210$

Предварительный анализ данных. Визуализация

$\infty \rightarrow 695020 \checkmark 0$
 $\infty \rightarrow 504020 - 0$
 $\infty \rightarrow 504020 + 0$
 $\infty \rightarrow 695020 - 0$
 $\infty \rightarrow 694020 - 0$
 $\infty \rightarrow 504020 - 0$
 $\infty \rightarrow 505020 \checkmark 0$
 $\infty \rightarrow 695020 - 0$
 $\infty \rightarrow 694020 - 0$

Предварительный анализ данных.

Визуализация

0	1	2	3	4	5	6	7	8	9
3286	3715	3360	3462	3209	3062	3296	3545	3268	3332

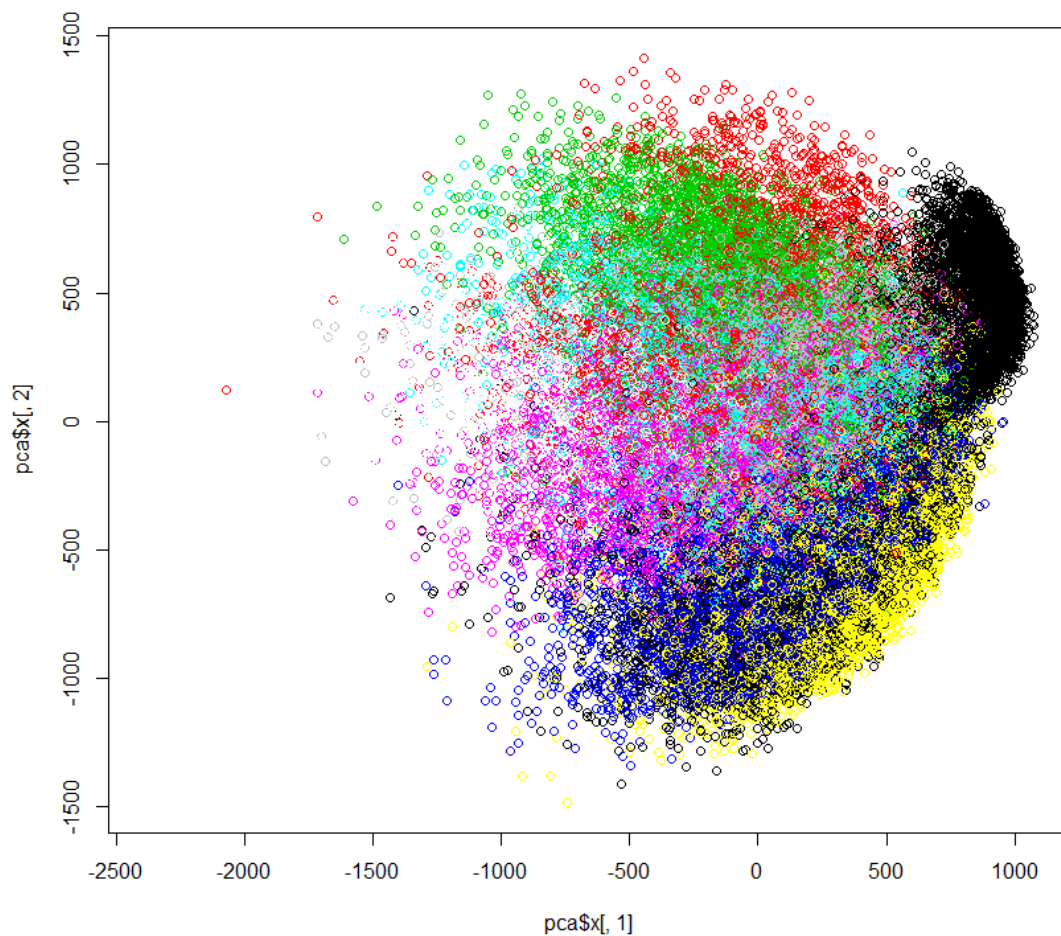


Уменьшение размерности данных для визуализации

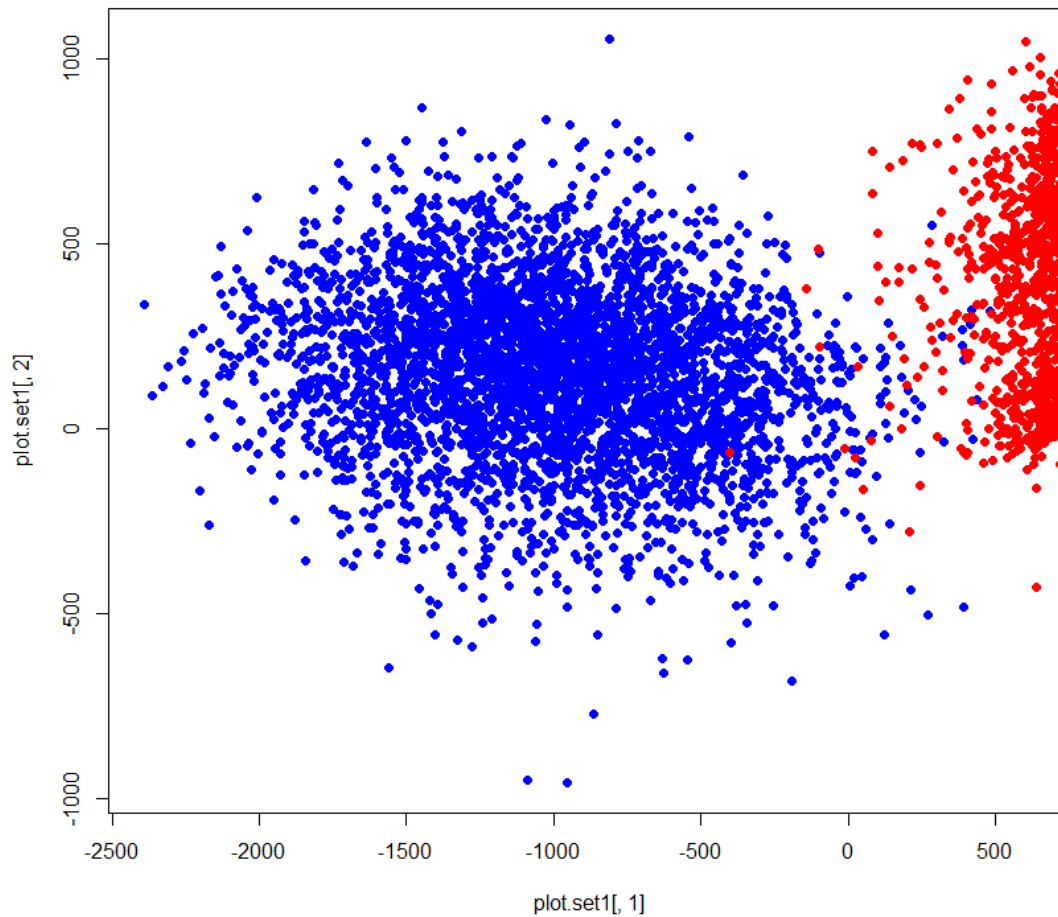
- ▶ Взаимное расположение классов на плоскости
 - ▶ длина одного вектора-признака $784 = 28 \times 28$ элементов
 - ▶ нужно оставить только две координаты, но максимально точно сохранить взаимное расположение векторов
 - ▶ методы уменьшения размерности данных
 - ▶ PCA, метод главных компонент
 - ▶ SVD, сингулярное разложение
 - ▶ MDS, многомерное шкалирование
- ▶ Признаки в 2D, 3D



Уменьшение размерности данных для визуализации



Уменьшение размерности данных для визуализации

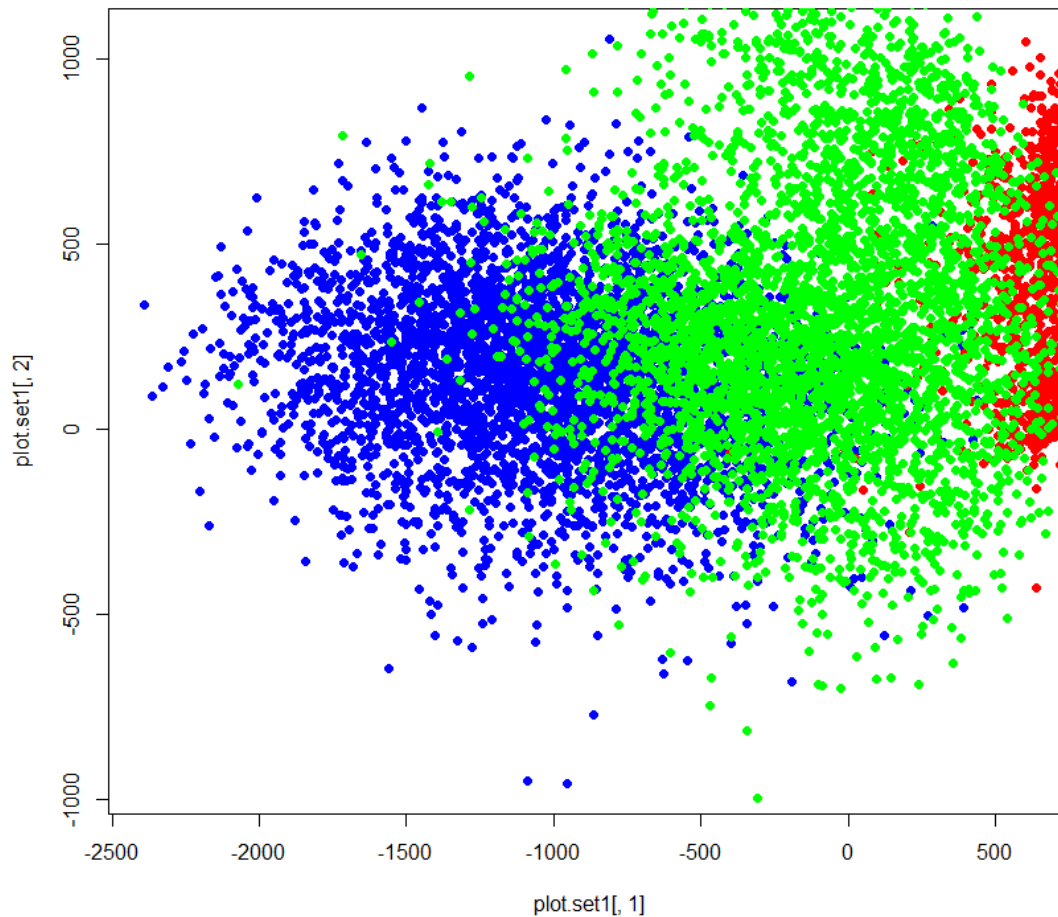


0 — синий цвет

1 — красный цвет



Уменьшение размерности данных для визуализации



0 — синий цвет
1 — красный цвет
2 — зелёный цвет



Классификация

- ▶ На тестовых данных с известными метками классов тренируется классификатор
 - ▶ метод k ближайших соседей
 - ▶ случайный лес
- ▶ Для неизвестных данных без меток классификатор принимает решение о принадлежности к одному из классов
 - ▶ метод k ближайших соседей, точность 96.338%
 - ▶ случайный лес, 96.645%
 - ▶ лучший результат для этих данных 99.471%



Уменьшение размерности данных для классификации

- ▶ Большие данные требуют много вычислительных ресурсов
 - ▶ процессорное время
 - ▶ оперативная память
- ▶ Матрица 42000×784 , файл 73 мегабайта, оперативная память во время классификации около 2 гигабайт
- ▶ Что делать
 - ▶ предоставить больше ресурсов (параллелизация, облачные вычисления)
 - ▶ преобразовать исходные данные с наименьшими потерями



Уменьшение размерности данных для классификации



Уменьшение размерности данных для классификации

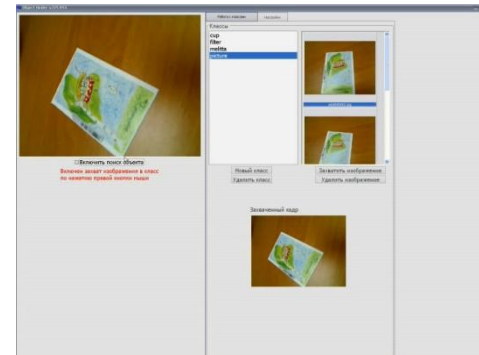
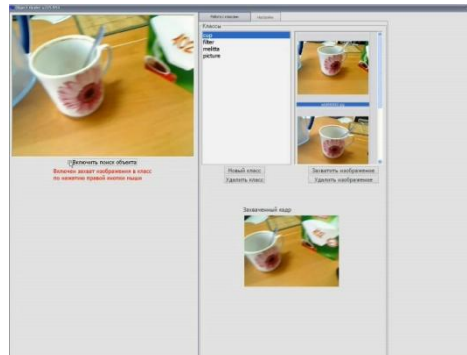
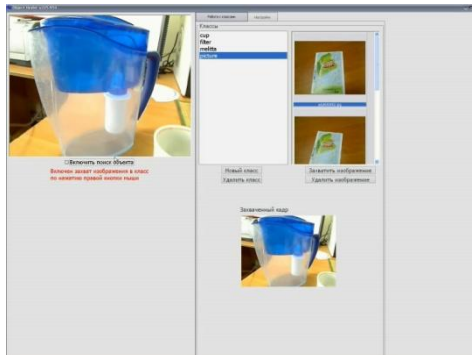
	10	154	200	784
К ближайших	92.233	96.527	96.527	96.338
Случайный лес	90.667	94.247	94.034	96.645

- Не всегда чем больше признаков, тем лучше
- Важно правильно выбрать признаки, классификатор и его параметры



Демонстрационное видео

- ▶ Процедура классификации для разных задач
 - ▶ извлечение признаков из изображения
 - ▶ преобразование признаков, выбор классификатора и подбор оптимальных параметров
 - ▶ использование обученного классификатора
 - ▶ распознавание рукописных цифр
 - ▶ детектирование новообразований на КТ лёгких
 - ▶ поиск объектов в комнате



Дополнительные материалы и ссылки

- ▶ Материалы этого доклада: презентация, данные, изображения, скрипты, видео-демонстрация
 - ▶ <https://github.com/ihar/BSU-Workshop-2012>
- ▶ Язык программирования для статистической обработки и программная среда вычислений R
 - ▶ <http://www.r-project.org/>
- ▶ База рукописных цифр и соревнование по их распознаванию
 - ▶ <http://www.kaggle.com/c/digit-recognizer>

