CSC411: Midterm Review

Xiaohui Zeng

February 7, 2019

Built from slides by James Lucas, Joe Wu and others

Agenda

1. A brief overview
2. Some sample questions

Basic ML Terminology

- Regression
- Overfitting
- Generalization
- Stochastic Gradient Descent (SGD)
- Classification
- Underfitting
- Regularization
- Bayes Optimal

Basic ML Terminology

- Training Data
- Validation Data
- Test Data

- Optimization
- 0-1 Loss
- Linear classifier

- Features
- Model

Some Questions

Question

1. Bagging improved performance by reducing __________

Some Questions

Question

1. Bagging improved performance by reducing variance

Some Questions

Question

1. Bagging improved performance by reducing variance
2. Given discrete random variables X and Y. The Information Gain in Y due to X is

$$IG(Y|X) = H(\_\_\_\_) - H(\_\_\_\_\_\_\_\_),$$

where H is the entropy

Some Questions

Question

1. Bagging improved performance by reducing variance
2. Given discrete random variables X and Y. The Information Gain in Y due to X is

$$IG(Y|X) = H(Y) - H(Y|X),$$

   where H is the entropy

Some Questions

Question 2

Take labelled data $(\mathbf{X}, \mathbf{y})$.

1. Why should you use a validation set?

Some Questions

Question 2

Take labelled data $(\mathbf{X}, \mathbf{y})$.

1. Why should you use a validation set?
2. How do you know if your model is overfitting?

Some Questions

Question 2

Take labelled data $(\mathbf{X}, \mathbf{y})$.

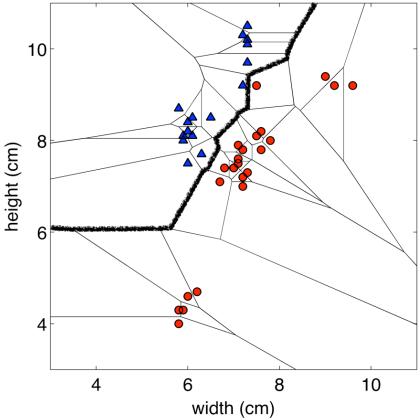
1. Why should you use a validation set?
2. How do you know if your model is overfitting?
3. How do you know if your model is underfitting?

ML Models

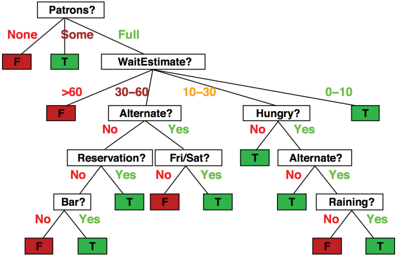
1. Nearest Neighbours
2. Decision Trees
3. Ensembles
4. Linear Regression
5. Logistic Regression
6. SVMs

Nearest Neighbours

1. Decision Boundaries
2. Choice of ‘k‘ vs. Generalization
3. Curse of dimensionality



Decision Trees



1. Entropy: $H(X)$, $H(Y|X)$
2. Information Gain
3. Decision Boundaries

Bayes Optimality

Starting with square error: $\mathbb{E}[(y-t)^2|\mathbf{x}] = \mathbb{E}[y^2 - 2yt + t^2|\mathbf{x}]$;

Bayes Optimality

Starting with square error: $\mathbb{E}[(y-t)^2|\mathbf{x}] = \mathbb{E}[y^2 - 2yt + t^2|\mathbf{x}]$;

1. choose a single value y^* based on $p(t|\mathbf{x})$:

$$(y - \mathbb{E}[t|\mathbf{x}])^2 + \mathrm{Var}(t|\mathbf{x})$$

Bayes Optimality

Starting with square error: $\mathbb{E}[(y-t)^2|\mathbf{x}] = \mathbb{E}[y^2 - 2yt + t^2|\mathbf{x}]$;

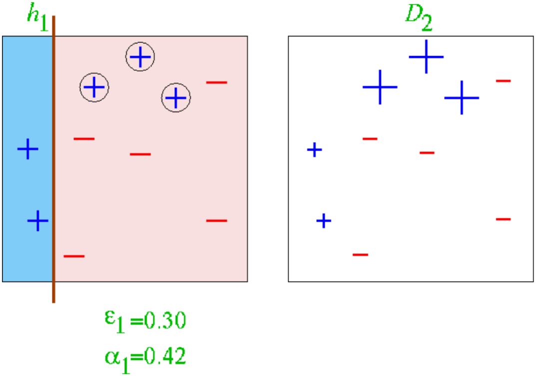
1. choose a single value y^* based on $p(t|\mathbf{x})$:

$$(y - \mathbb{E}[t|\mathbf{x}])^2 + \mathrm{Var}(t|\mathbf{x})$$

2. treat y as a random variable:
   - bias term
   - variance term
   - Bayes error

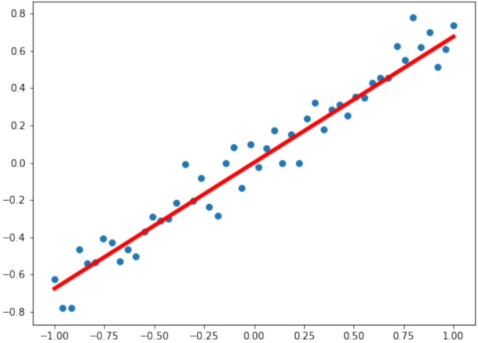
Ensembles

1. Bagging/bootstrap aggregation
2. Boosting
   - decision stumps:

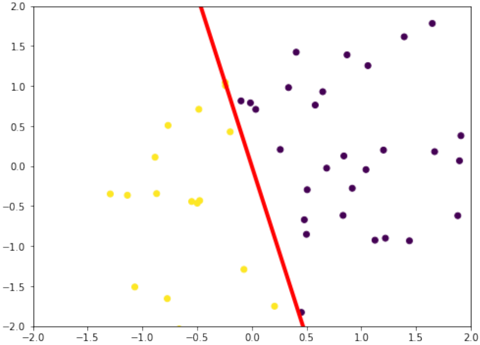


Linear Regression

1. Loss function
2. Direct solution
3. (Stochastic) Gradient Descent
4. Regularization
   - L_1 vs L_2 norm



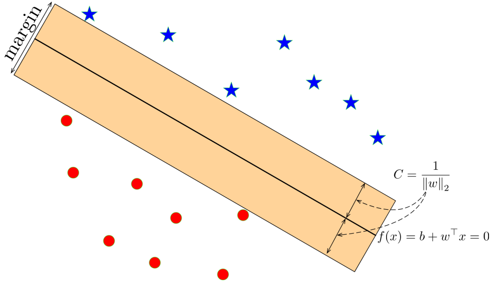
Logistic Regression



1. Loss functions
   - 0-1 loss?
   - l2 loss?
   - cross-entropy loss?
2. Binary vs. Multi-class
3. Decision Boundaries
   - $\hat{p} = \frac{1}{1+e^{-\theta x}}$

SVMs

1. Hinge loss
2. Margins



Sample Question 1

First, we use a linear regression method to model this data.
To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set.
Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors? (No explanation required)

- Mean Training Error: A. Increase; B. Decrease
- Mean Testing Error: A. Increase; B. Decrease

Q1 Solution

- The training error tends to increase. As more examples have to be fitted, it becomes harder to 'hit', or even come close, to all of them.

Q1 Solution

- The training error tends to increase. As more examples have to be fitted, it becomes harder to 'hit', or even come close, to all of them.
- The test error tends to decrease. As we take into account more examples when training, we have more information, and can come up with a model that better resembles the true behavior. More training examples lead to better generalization.

Sample Question 2

If variables X and Y are independent, is $I(X|Y) = 0$? If yes, prove it. If no, give a counter example.

Q2 Solution

Recall that

- two random variables X and Y are independent if for all $x \in \text{Values}(X)$ and all $y \in \text{Values}(Y)$, $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- $H(X) = -\sum_x P(x)\log_2 P(x)$

Q2 Solution

$$
\begin{aligned}
I(X|Y) &= H(X) - H(X|Y) && (1)\\
&= -\sum_x P(x)\log_2 P(x) - \big(-\sum_y \sum_x P(x,y)\log_2 P(x|y)\big) && (2)\\
&= -\sum_x P(x)\log_2 P(x) - \big(-\sum_y P(y) \sum_x P(x)\log_2 P(x)\big) && (3)\\
&= -\sum_x P(x)\log_2 P(x) - \big(-\sum_x P(x)\log_2 P(x)\big) && (4)\\
&= 0 && (5)
\end{aligned}
$$

Sample Question 3

Given input $\mathbf{x} \in \mathbb{R}^D$ and target $t \in \mathbb{R}$, consider a linear model of the form: $y(x, \mathbf{w}) = \sum_{i=1}^{D} w_i x_i$.. Now suppose a noisy pertubation ϵ_i is added independently to each of the input variables x_i. i.e., $\hat{x}_i = x_i + \epsilon_i$, assume

- $\mathbb{E}[\epsilon_i] = 0$
- for $i \neq j$: $\mathbb{E}[\epsilon_i \epsilon_j] = 0$
- $\mathbb{E}[\epsilon_i^2] = \lambda$

We define the following objective that tries to be robust to noise

$$\mathbf{w}^* = \arg\min \mathbb{E}_\epsilon[(\mathbf{w}^T \hat{\mathbf{x}} - t)^2]. \tag{6}$$

Show that it is equivalent to minimizing L_2 regularized linear regression, *i.e.*,:

$$\mathbf{w}^* = \arg\min[(\mathbf{w}^T \mathbf{x} - t_n)^2 + \lambda ||\mathbf{w}||^2].$$

Q3 Solution

Let

$$\hat{y} = \sum_{i=1}^{D} w_i(x_i + \epsilon_i) = y + \sum_{i=1}^{D} w_i \epsilon_i,$$

where $y = \mathbf{w}^T \mathbf{x} = \sum_{i=1}^{D} w_i x_i$.

Q3 Solution

Then we start with

$$\mathbf{w}^* = \arg\min \mathbb{E}_\epsilon[(\mathbf{w}^T \hat{\mathbf{x}} - t)^2] = \arg\min \mathbb{E}_\epsilon[(\hat{y} - t)^2],$$

the inner term

$$\begin{aligned}
(\hat{y} - t)^2 &= \hat{y}^2 - 2\hat{y}t + t^2 &(7)\\
&= (y + \sum_{i=1}^{D} w_i\epsilon_i)^2 - 2t(y + \sum_{i=1}^{D} w_i\epsilon_i) + t^2 &(8)\\
&= y^2 + 2y\sum_{i=1}^{D} w_i\epsilon_i + (\sum_{i=1}^{D} w_i\epsilon_i)^2 - 2ty - 2t\sum_{i=1}^{D} w_i\epsilon_i + t^2 &(9)
\end{aligned}$$

then we take the expectation under the distribution of ϵ

Q3 Solution

That is,

$$\mathbb{E}_\epsilon\big[y^2 + 2y\sum_{i=1}^{D} w_i\epsilon_i + (\sum_{i=1}^{D} w_i\epsilon_i)^2 - 2ty - 2t\sum_{i=1}^{D} w_i\epsilon_i + t^2\big]$$

we have the second and the fifth term equal to zero since $\mathbb{E}[\epsilon_i] = 0$, while the third term become $\mathbb{E}_\epsilon[(\sum_{i=1}^{D} w_i\epsilon_i)^2] = \lambda\sum_{i=1}^{D} w_i^2$.
Finally we see

$$\mathbb{E}_\epsilon\big[y^2 + \lambda\sum_{i=1}^{D} w_i^2 - 2ty + t^2\big] = \mathbb{E}_\epsilon[(y-t)^2 + \lambda\sum_{i=1}^{D} w_i^2].$$