CSC411: Midterm Review

James Lucas

October 11, 2018

Built from slides by Joe Wu and others

Agenda

1. A brief overview
2. Some sample questions

Basic ML Terminology

- Regression
- Overfitting
- Generalization
- Maximum Likelihood (MLE)
- Stochastic Gradient Descent (SGD)
- Classification
- Underfitting
- Regularization
- Maximum a posteriori (MAP)
- Bayes Optimal

Basic ML Terminology

- Model
- Linear classifier
- Training Data
- Discriminative approach
- Optimization
- 0-1 Loss
- Validation Data
- Generative approach
- Convexity
- Features
- Test Data
- Bayesian approach

Some Questions

Question 1

Given a discriminative model with parameters θ and training data $(\mathbf{X}, \mathbf{y})$.

1. The likelihood is _____
2. MAP maximizes _____

Question 2

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

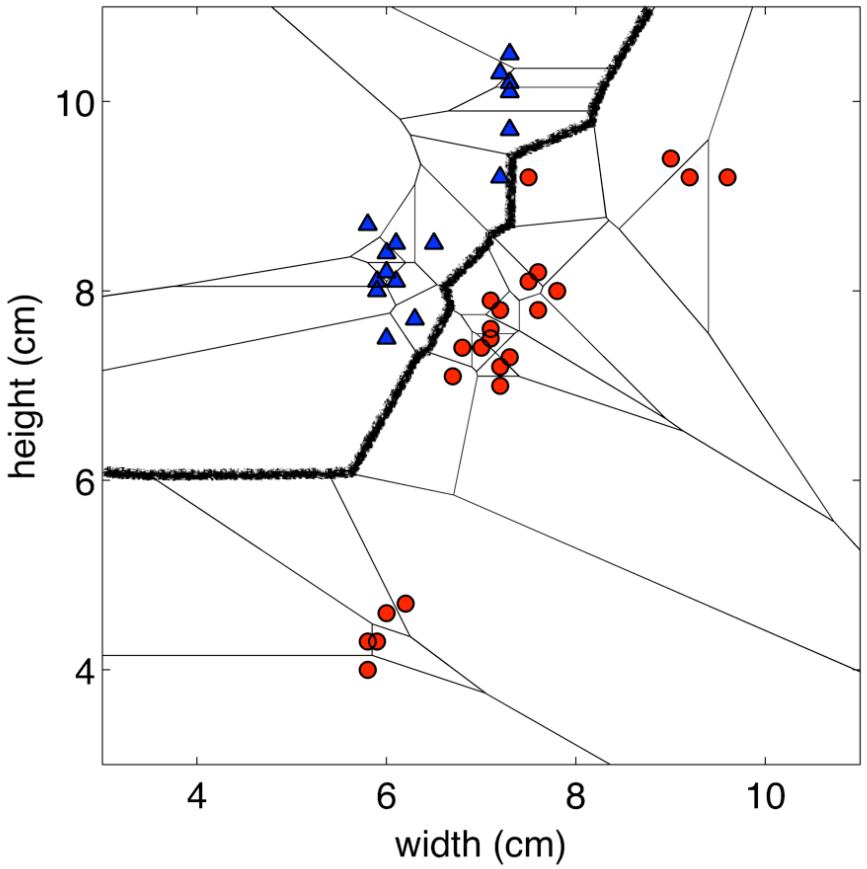
1. Why should you use a validation set?
2. How do you know if your model is overfitting?
3. How do you know if your model is underfitting?

ML Models

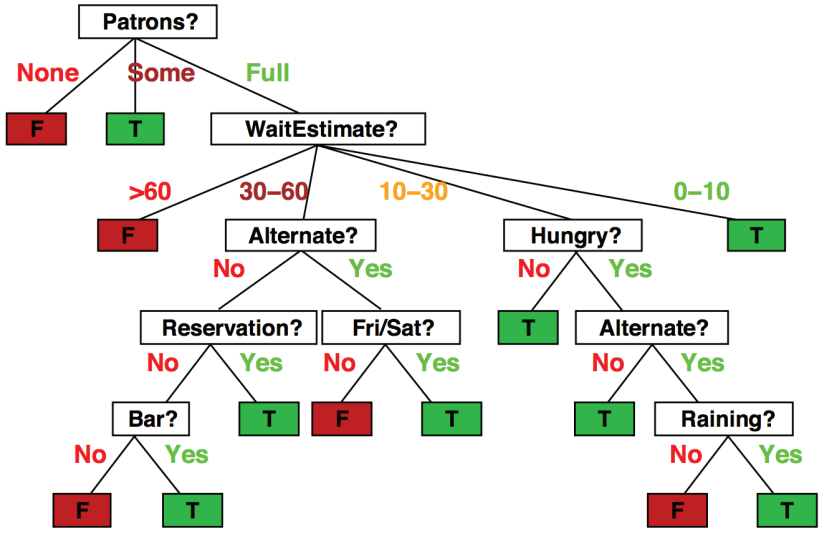
1. Nearest Neighbours
2. Decision Trees
3. Ensembles
4. Linear Regression
5. Logistic Regression
6. SVMs

Nearest Neighbours

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



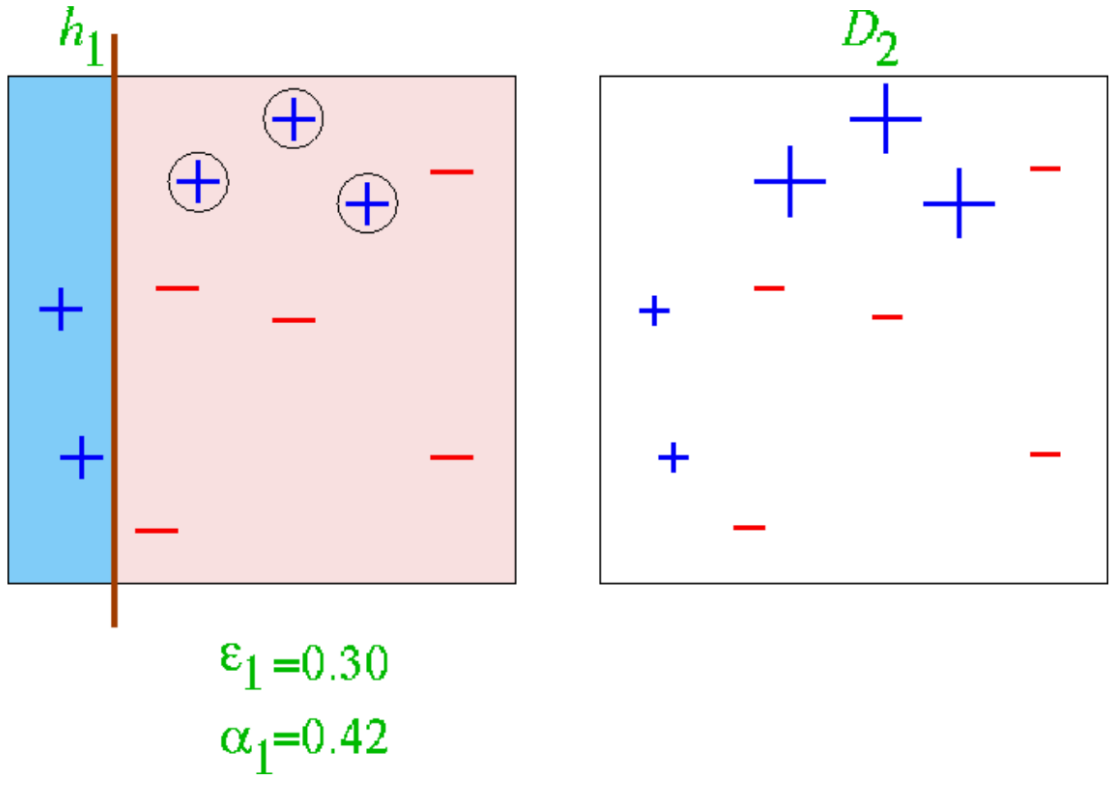
Decision Trees



1. Entropy/Information Gain
2. Decision Boundaries

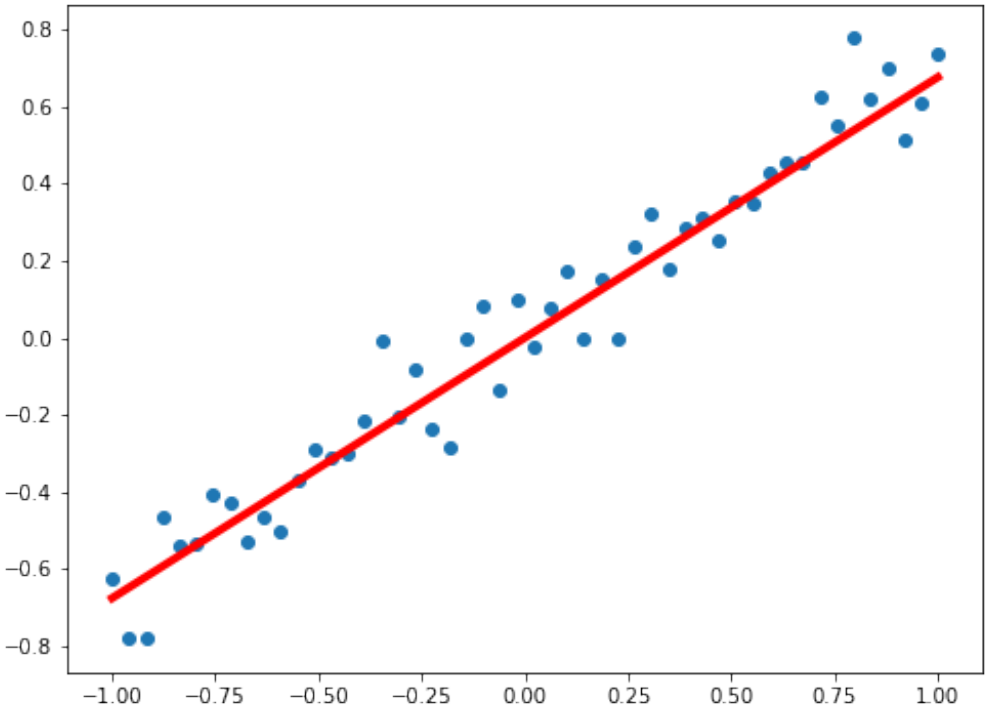
Ensembles

1. Bagging
2. Boosting

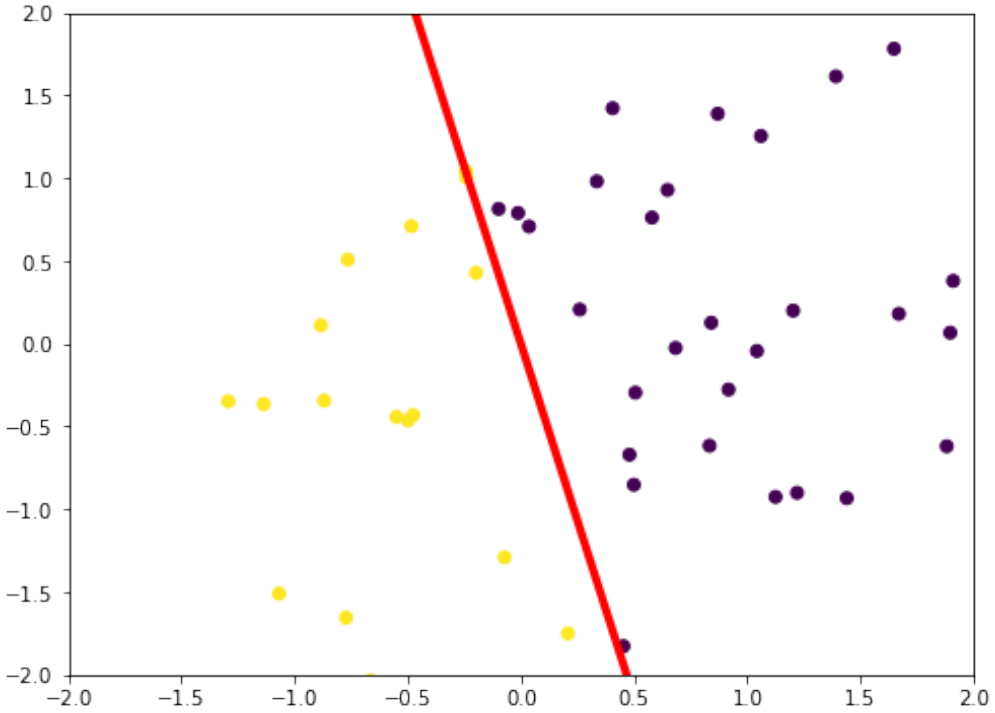


Linear Regression

1. Loss function
2. Direct solution
3. (Stochastic) Gradient Descent
4. Regularization



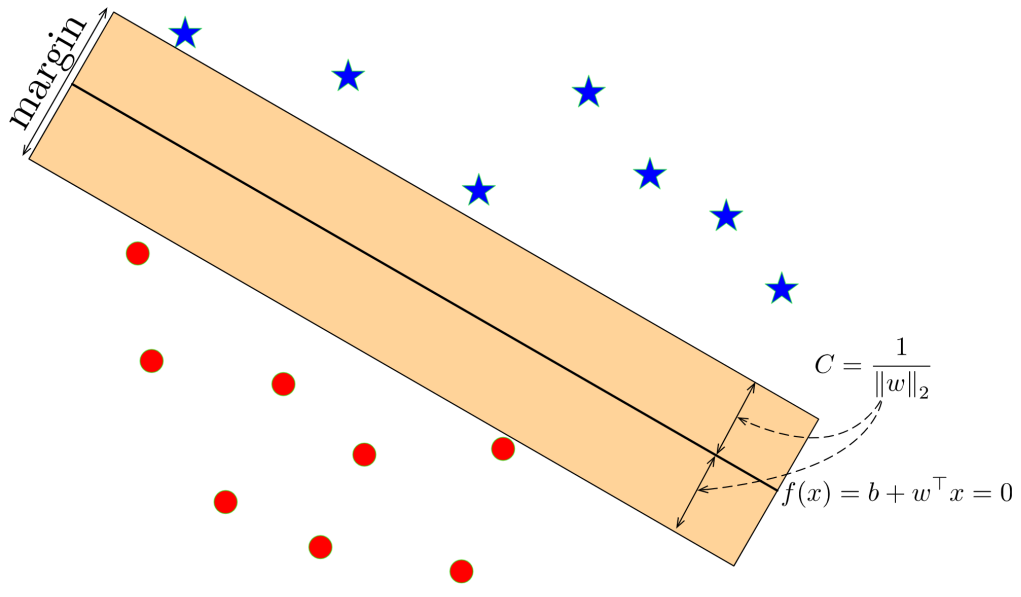
Logistic Regression



1. Loss functions
2. Binary vs. Multi-class
3. Decision Boundaries

SVMs

1. Hinge loss
2. Margins



Sample Question 1

Assume we are preprocessing our data using an **invertible** linear transformation on the features of our training data. The transformation can either be some orthogonal (i.e. rotations) matrix or some diagonal matrix.
Say if this can have any effect on the performance of the following algorithms, and explain in no more than two sentences.

- Orthogonal preprocessing on decision tree classification.
- Diagonal preprocessing on decision tree classification.
- Orthogonal preprocessing on nearest neighbor classification.
- Diagonal preprocessing on nearest neighbor classification.

Q1 Solution

- Orthogonal preprocessing on decision tree classification.
  Will have an effect. Rotation changes the axis.
- Diagonal preprocessing on decision tree classification.
  Will not have an effect. Rescaling along axis will shift split criteria but wont change decision.
- Orthogonal preprocessing on nearest neighbor classification.
  Will not have an effect. Orthogonal linear transformations will preserve distances.
- Diagonal preprocessing on nearest neighbor classification.
  Will have an effect. Will change distances between data points.

Sample Question 2

Given input $\mathbf{x} \in \mathbb{R}^d$ and target $y \in \mathbb{R}$, define $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\epsilon}$ to be a noisy pertubation of $\mathbf{x}$ where we 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

$$\mathbf{w}^* = \arg\min \mathbb{E}_\epsilon[(\mathbf{w}^T \hat{\mathbf{x}} - y)^2] \tag{1}$$

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

$$\mathbf{w}^* = \arg\min \left[(\mathbf{w}^T \mathbf{x} - y)^2 + \lambda ||\mathbf{w}||^2\right] \tag{2}$$

Q2 Solution

We can write the inner term as,

$$
\begin{aligned}
(\mathbf{w}^T\hat{\mathbf{x}} - y)^2 &= (\mathbf{w}^T\mathbf{x} + \mathbf{w}^T\boldsymbol{\epsilon} - y)^2 && (3)\\
&= (\mathbf{w}^T\mathbf{x} - y)^2 + 2\mathbf{w}^T\boldsymbol{\epsilon}(\mathbf{w}^T\mathbf{x} - y) + (\mathbf{w}^T\boldsymbol{\epsilon})^2 && (4)\\
&= (\mathbf{w}^T\mathbf{x} - y)^2 + 2\mathbf{w}^T\boldsymbol{\epsilon}(\mathbf{w}^T\mathbf{x} - y) + (\mathbf{w}^T\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\mathbf{w}) && (5)
\end{aligned}
$$

Under the expectation the second term will be zero as it is a linear combination of the elements of $\boldsymbol{\epsilon}$. The final term will be the quadratic form of $\mathbf{w}$ with the covariance of $\boldsymbol{\epsilon}$. The covariance is simply λI. Thus we are minimizing,

$$(\mathbf{w}^T\mathbf{x} - y)^2 + \lambda||\mathbf{w}||^2$$

which is exactly the objective of L2-regularized linear regression.