Module 4 - Linear Classifiers : Overfitting & Regularization

Training the classifier: ¶

- The dataset -> (labelled -> [(sentence1, +), (sentence2, -), ...]) is provided.
- The dataset is divided into **Training set** and **Validation set**.
- The training set is fed into the learn classifier and the algorithm is built, model is trained.
- The learnt model is run on the validation set and evaluate the classifier.

Classification Error

- 'Classification error' - It is the measure of the classifiers performance.
- If the 'true label' and 'predicted label' are the same then it is right else it is wrong.

Classification Error

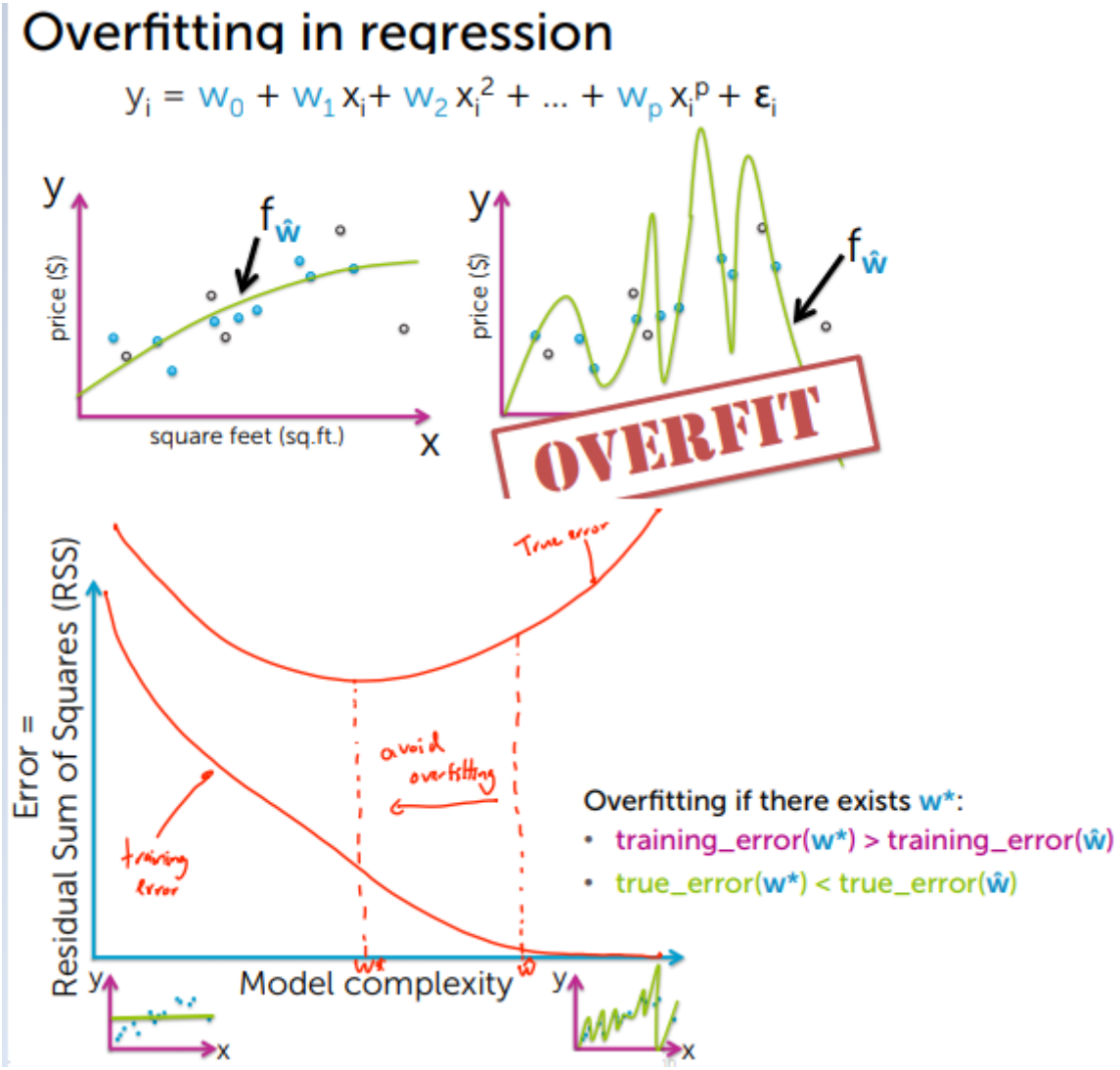
- Error measures the fraction of mistakes.
- **error = #mistakes / total number of datapoints**.
- Best possible value = 0.0.

Accuracy

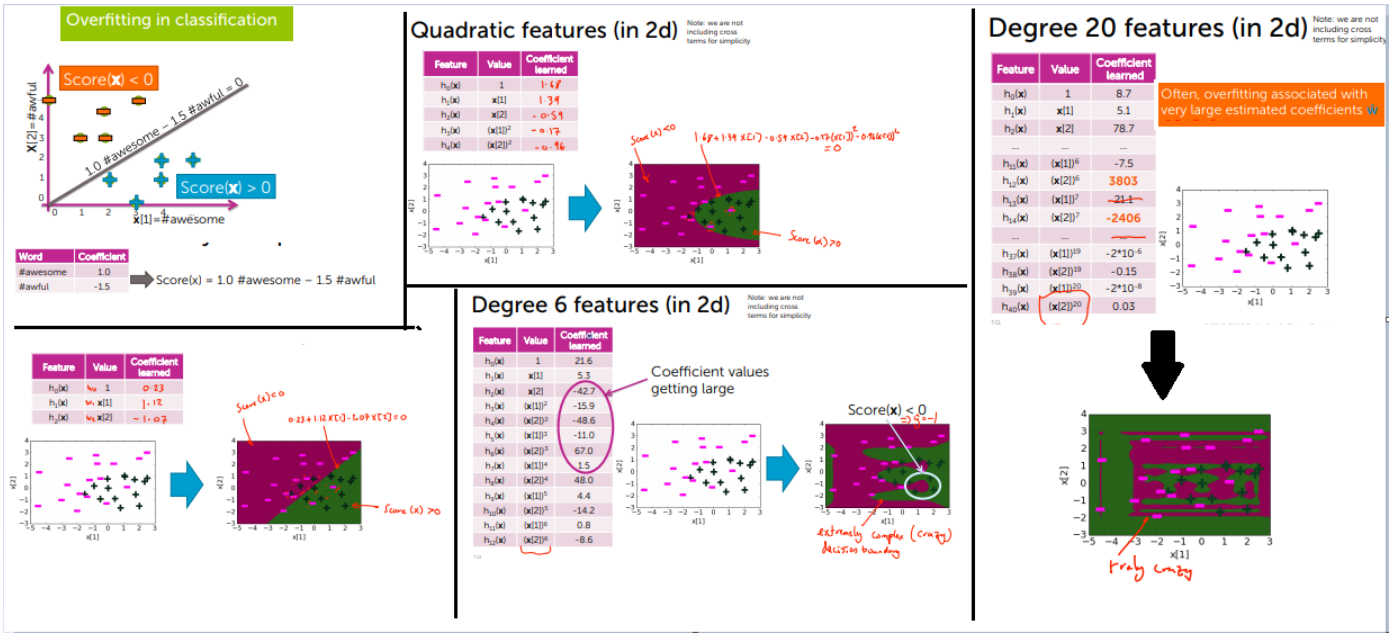
- Fraction of correct predictions.
- **accuracy = #correct / total number of datapoints**.
- Best possible value - 1.0.

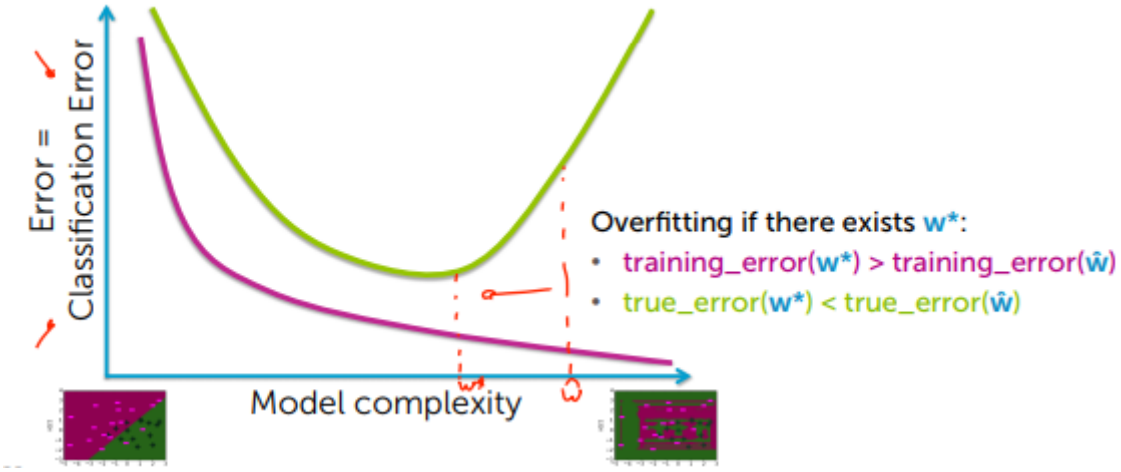
Overfitting in regression

- Here as the model complexity increases, the model fits the training data very well, but might not be generalized leading to **overfitting**.



Overfitting in Classification





Overfitting in Classifiers -> Overconfident Predictions

- In the 'Logistic Regression Model' the score -> w(transpose) * features -> extendes [-infinity, infinity];
- This is reduce to their probability funtion of the output given input using the sigmoin/link function in the range [0,1] -> where the output ranges y [-1, +1];

The subtle consequences of overfitting in logistic regression

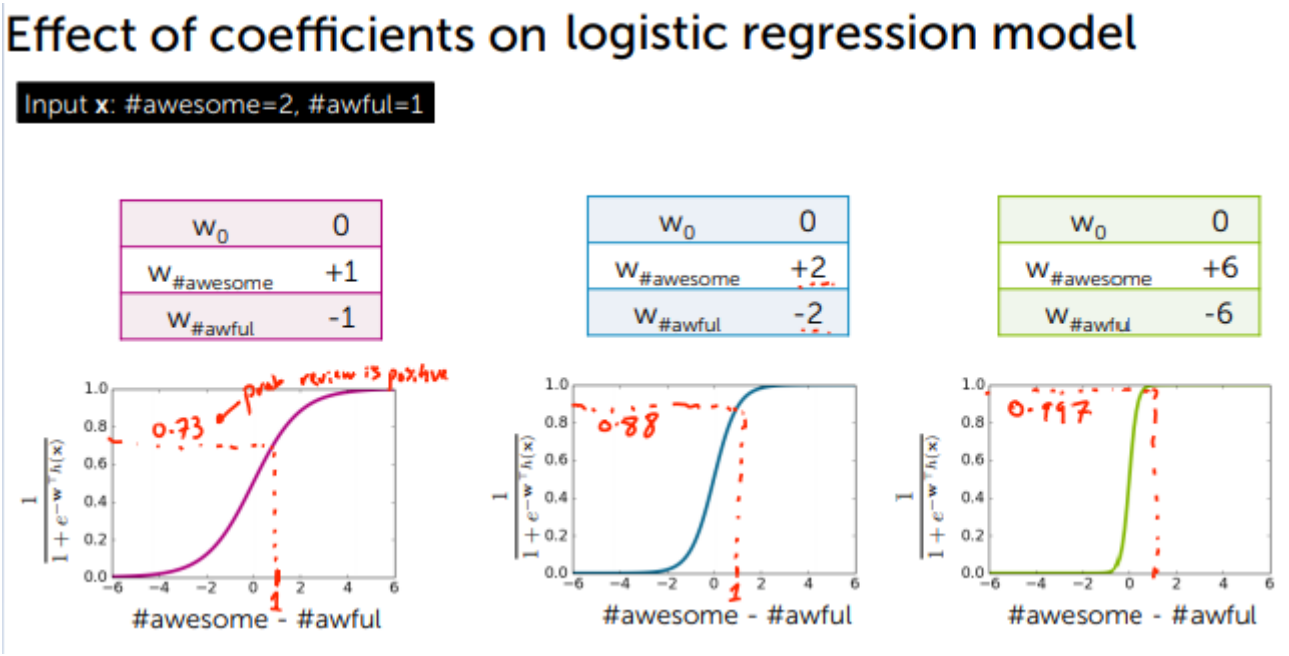
- Overfitting leads to 'Large coefficient values'.

```
    - wT * h(xi) -> either very positive / very negative-> which makes the si
gmoid(wT * h(xi)) -> go to 0 or 1.
```

- The model becomes extremely overconfident of predictions.
- Below 3 cases are listed for the same

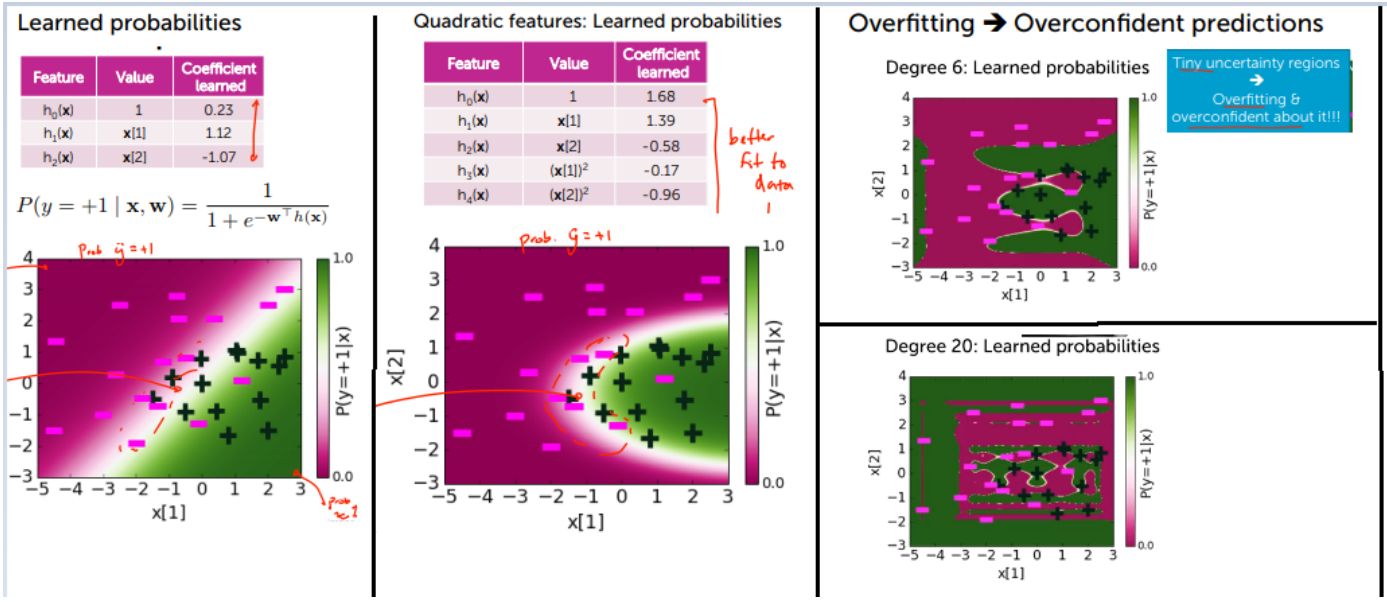
```
* Input : #awesome = 2, #awful = 1
* X-axis = #awesome - #awful = 2 - 1 = 1
```

```
- Coefficients are intercept, weight -awesome, weight -awful
* Y-axis =  Score = 1 / (1 + e^-(wT * h(xi));
1. Coefficients -> 0, +1, -1  -> 0.73
2. Coefficients -> 0, +2, -2  -> 0.88
3. Coefficients -> 0, +6, -6  -> 0.997
```



Visualizing the dataset to identify Overfitting - Overconfident predictions

- As model complexity increases, the decision boundary between the classification data becomes very this indicates overfitting.
- The separation between the data must be narrow but not very wide or extremely tiny. **In case of tiny it is an indication of overfitting**.



Another perspective on overfitting logistic regression (ADVANCED)

Linearly-separable data: A dataset that can be classified into categories. A line can be drawn to segregate the data.

- Data is linearly separable if: There are coefficients w-hat such that:

```
- For all positive training data: Score(x) = wT h(x) >0
- For all negative training data: Score(x) = wT h(x) <0
```

- **training_error(w-hat) = 0. For linearly-separable data the training_error = 0.** *This could be a situation of overfitting - especially w.r.t complex models.*
- **Note 1 : If there are D features, linear separability happens in a D-dimensional space.**

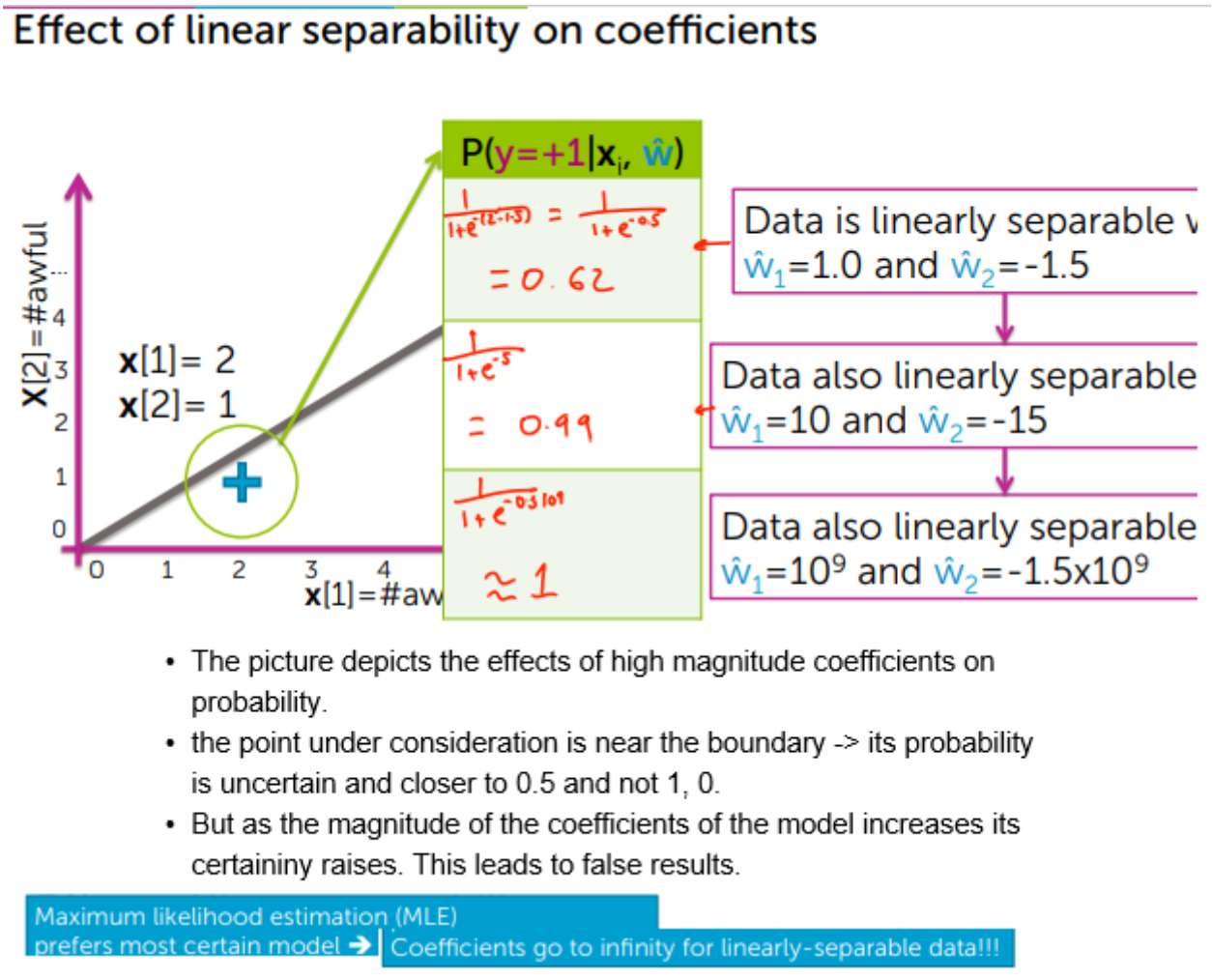
- **Note 2 : If you have enough features, data are (almost) always linearly separable.** *.
- Polynomial to the degree 50,100, 180, etc - data is gonna become linearly separated - lead to Problematic case.

Effects of linearly separability on coefficients

- Consider a plane that separates the data (positive and negative);
  - Plane -> 1.0 #awesome - 1.5 #awful = 0
  - Multiplying * 10 -> 10 #awesome - 15 #awful = 0
  - Multiplying * 10^9 -> 10x10^9 #awesome - 15x10^9 #awful = 0
    - In the above case -> although the values of the coefficients is increasing, the plane separating the positive and negative boundaries is still the same. Hence the prediction are not right if its a result of increase in magnitude of the coefficients.

Issue : MLE (Maximum likelihood estimation) - prefers most certain models, but here in case of overfit models / linearly-separable data the coefficients go to infinity, increasing the certainty of the prediction -> This is problem.

- The picture depicts the effects of high magnitude coefficients on probability.
- the point under consideration is near the boundary -> its probability is uncertain and closer to 0.5 and not 1, 0.
- But as the magnitude of the coefficients of the model increases its certaininy raises. This leads to false results.



Overfitting in logistic regression

- Learning tries to find decision boundary that seprates data - 'Overly complex boundary'.
- If data is linearly separable -> coefficients go to infinity.

L2 regularized logistic regression

Penalizing large coefficients to mitigate overfitting

- Quality Metric is modified to handle - large coefficinets and prevent overfitting.

Desired total cost format

- Want to balance:
    1. How well the function fits the data -> large.
    2. Magnitude of the coefficients -> small.
- Total quality = measure of fit - measure of magnitude of coefficients.
    - Measure of fit -> data likelihood -> large # = good fit for training data.
    - Measure of magnitude of coefficients -> large # = overfit.

Part 1 : Maximum likelihood estimate (MLE):

- Measure of fit = Data likelihood
    - Choose coefficients of w that maximize likelihood.

$$\ell(\mathbf{w}) = \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

    - Typically, use log of likelihood function (simplifies math and has better gradient/convergence properties.)

$$\ell\ell(\mathbf{w}) = \ln \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

natural log

Data likelihood is supposed to be as big as possible.

Part 2 : Measure of magnitude of logistic regression coefficients:

- Sum of squares (L2 norm) -> penalize highly positive and highly negative numbers in the same way.
- Sum of absolute value (L1 norm) -> provides sparse solutions.
- Both the metrics penalize large coefficients.

L2 regularized logistic regression

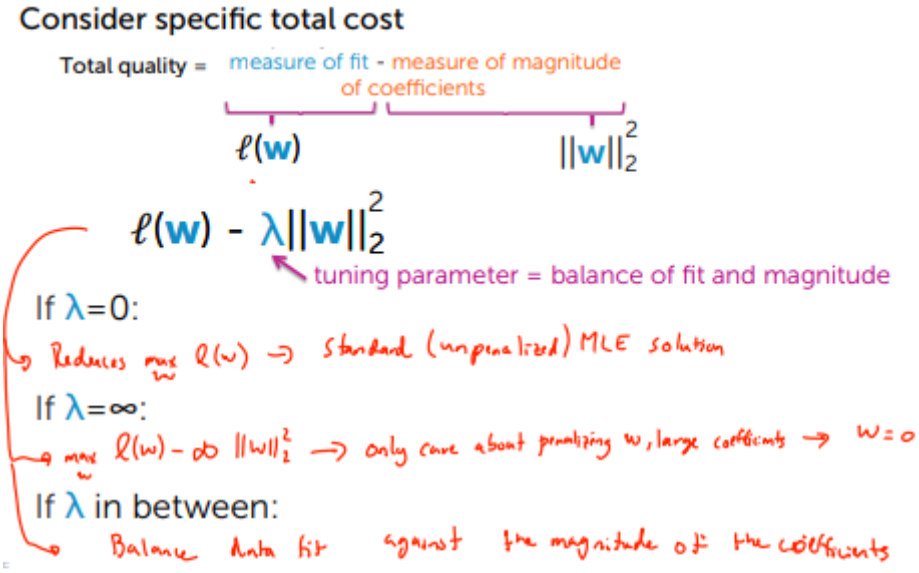
- The mechanism of finding **lambda**/ tuning parameter that balances between the model fit and the coefficient magnitude is **L2 regularized logistic regression**. (In regression case -> term - Ridge Regression);

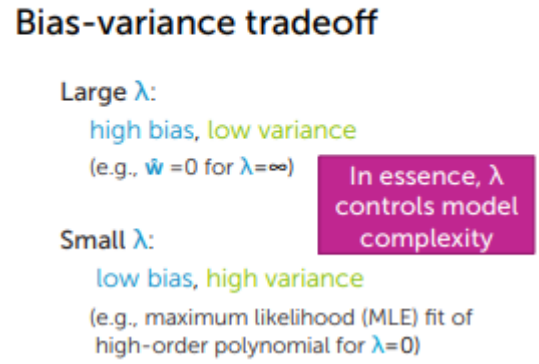
Picking lambda:

- Validation Set (for large dataset).
- Cross-validation (for smaller datasets).

Bias-Variance tradeoff

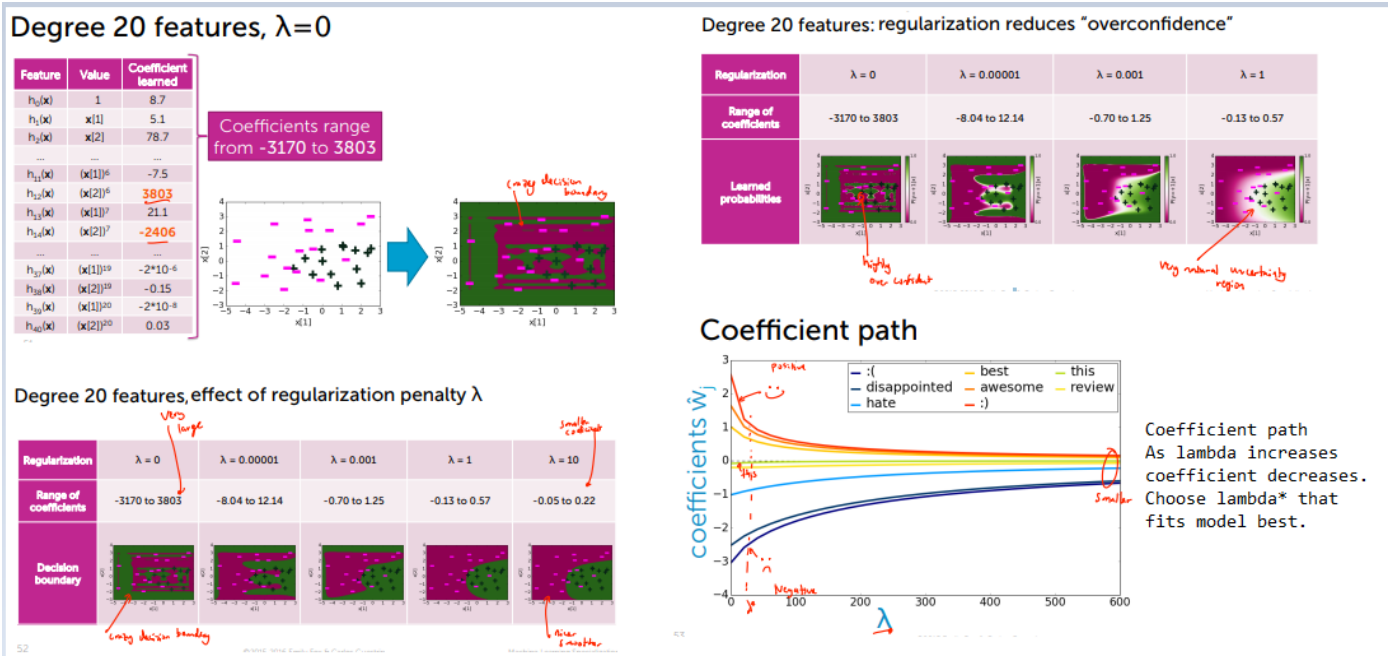
- Lambda controls the moel complexity.
    1. Large lambda : high bias, low variance. (w-hat = 0, lambda = infinity);
    2. Small lambda : low bias, high variance. (MLE for higher order polynomial, lambda = 0.





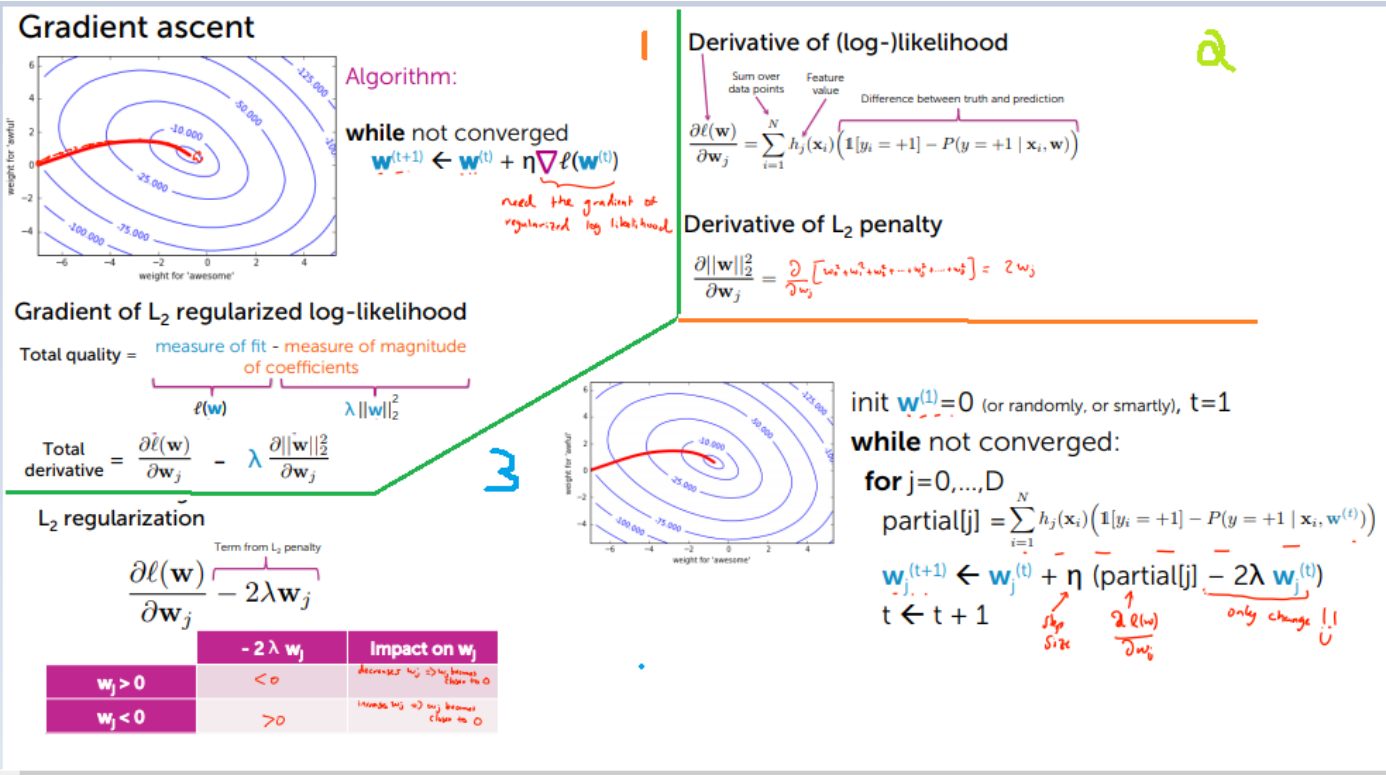
L2 regularization address overfitting issues.

- Choosing appropriate lambda value with higher order complex models
    1. Provides a better decision boundary between the data.
    2. The overconfidence predictions is reduced as a natural uncertainity region is obtained.



Learning L2 regularized logistic regression with gradient ascent

- Algorithm to optimize to get w-hat.
- Employ gradient ascent algorithm to find the w-hat.



Sparse logistic regression with L1 regularization

- This provides Efficiency and Interpretability.

Efficiency:

- If size(w-hat) = 100B, each prediction is expensive.
- If w-hat is sparse, computation only depends on the # non-zeros.

```
y-hat(i) = sign (Sum(wj!=0) w-hatj * hj(xi))
```

Interpretabiliy:

- Can decipher the features truly relevant for the prediction.

Sparse Logistic Regression

- Total quality = measure of fit - measure of magnitude of coefficients
- Total quality - l(w) - ||w||1
- L1 norm leads to sparse solutions.

L1 regularized logistic regression

- Lambda is a tuning parameter that provides a balance between the model fit and sparsity.

Coefficient path - L1 penalty

- The coefficients that contribute to the model become 0 eventually with increasing lamba. While those hardly contributing to the model become 0 initially.

Sparse logistic regression

$$\hat{y}_i = sign\left(\sum_{\hat{w}_j \neq 0} \hat{w}_j h_j(x_i)\right)$$

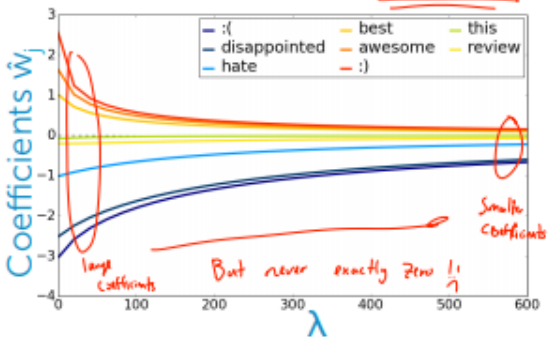
Total quality =
measure of fit - measure of magnitude of coefficients

$\ell(w)$ $\quad$ $||w||_1 = |w_0| + \ldots + |w_D|$

L_1 regularized logistic regression

Leads to sparse solutions!

Regularization path – L_2 penalty



L_1 regularized logistic regression

$$\ell(w) - \lambda||w||_1$$

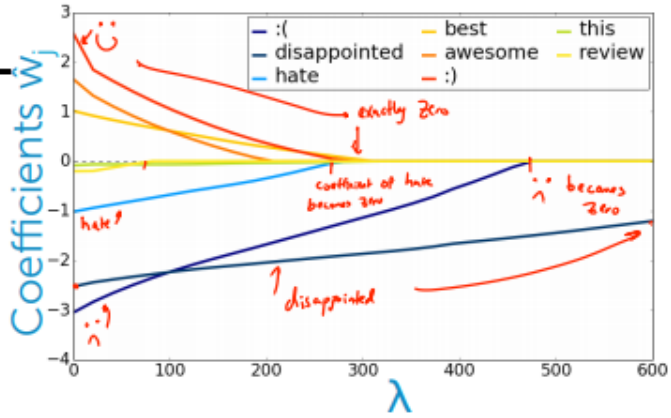
tuning parameter = balance of fit and sparsity

If λ=0:
- No regularization → standard MLE solution

If λ=∞:
- all weight is on regularization → $\hat{w}$ = 0

If λ in between:
- Sparse solutions: Some $\hat{w}_j \neq 0$, many other $\hat{w}_j = 0$

Regularization path – L_1 penalty



Quiz

1. Consider four classifiers, whose classification performance is given by the following table:

	Classification error on training set	Classification error on validation set
Classifier 1	0.2	0.6
Classifier 2	0.8	0.6
Classifier 3	0.2	0.2
Classifier 4	0.5	0.4

Which of the four classifiers is most likely overfit?

- (●) Classifier 1
- ( ) Classifier 2
- ( ) Classifier 3
- ( ) Classifier 4

---

2. Suppose a classifier classifies 23100 examples correctly and 1900 examples incorrectly. Compute error by hand. Round your answer to 3 decimal places.

0.076

---

3. (True/False) Accuracy and error measured on the same dataset always sum to 1.

- (●) True
- ( ) False

---

4. Which of the following is NOT a correct description of complex models?

- ( ) Complex models accommodate many features.
- ( ) Complex models tend to produce lower training error than simple models.
- (●) Complex models tend to generalize better than simple models.
- ( ) Complex models tend to exhibit high variance in response to perturbation in the training data.
- ( ) Complex models tend to exhibit low bias, capturing many patterns in the training data that simple models may have missed.

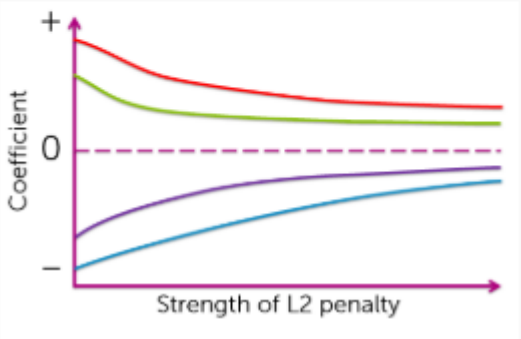
---

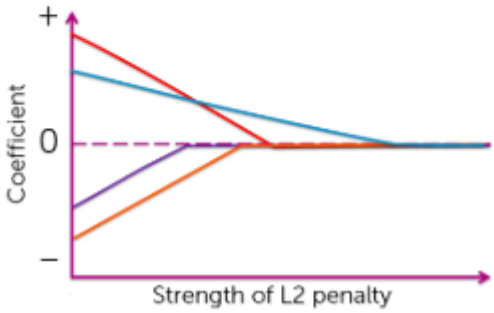
5. Which of the following is a symptom of overfitting in the context of logistic regression? Select all that apply.

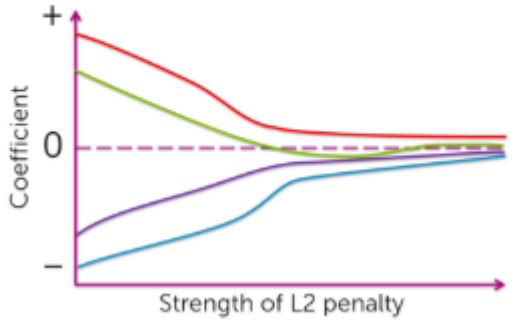
- [■] Large estimated coefficients
- [ ] Good generalization to previously unseen data
- [ ] Simple decision boundary
- [■] Complex decision boundary
- [■] Overconfident predictions of class probabilities

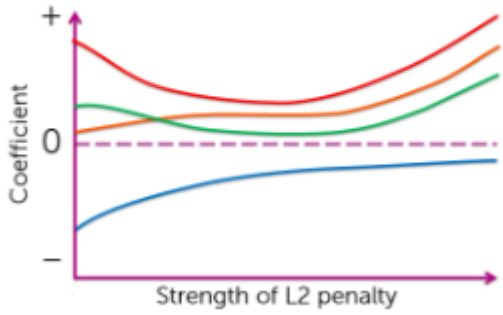
6. Suppose we perform L2 regularized logistic regression to fit a sentiment classifier. Which of the following plots does NOT describe a possible coefficient path? Choose all that apply.

Note. Assume that the algorithm runs for a wide range of L2 penalty values and each coefficient plot is zoomed out enough to capture all long-term trends.



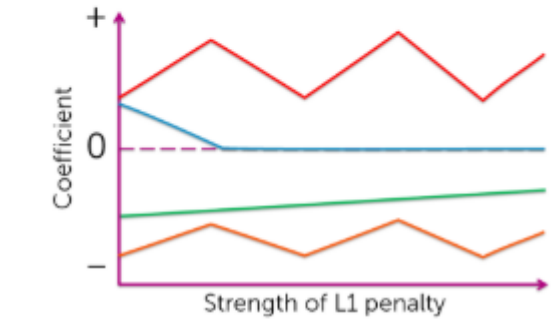


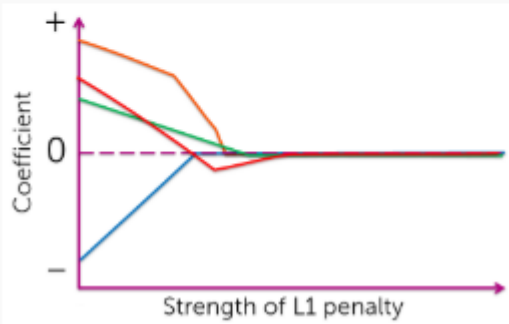


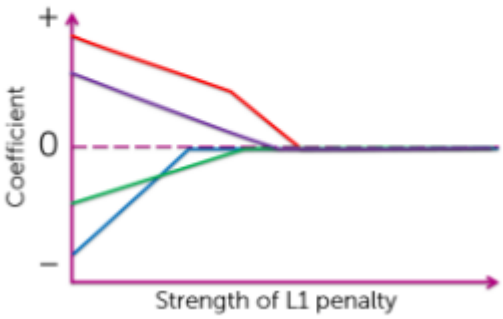


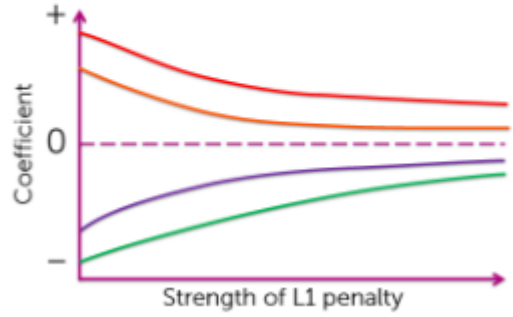
7. Suppose we perform L1 regularized logistic regression to fit a sentiment classifier. Which of the following plots does NOT describe a possible coefficient path? Choose all that apply.

Note. Assume that the algorithm runs for a wide range of L1 penalty values and each coefficient plot is zoomed out enough to capture all long-term trends.









8. In the context of L2 regularized logistic regression, which of the following occurs as we increase the L2 penalty λ? Choose all that apply.

- [x] The L2 norm of the set of coefficients gets smaller
- [ ] Region of uncertainty becomes narrower, i.e., the classifier makes predictions with higher confidence.
- [x] Decision boundary becomes less complex
- [ ] Training error decreases
- [x] The classifier has lower variance
- [ ] Some features are excluded from the classifier

In [ ]: