CS 181 LECTURE 2/1

Scribe Notes

Contents

1 Linear Regression

Linear regression, while powerful, can be limiting in scenarios where datasets are not linearly separable. The question arises: how can we extend our modeling capabilities?

1.1 Basis Expansion

One approach is to employ basis expansion, allowing us to transform the data and break free from linear limitations. Instead of relying on linear separability, we introduce a set of basis functions denoted as ϕ. The transformation is applied to the data, and we work with the linear combination of these basis functions.

1.1.1 Transformed Data Representation

Let $\mathbf{X}$ represent our original data. We now use $\Phi(\mathbf{X})$ to denote the transformed data, where Φ is a function applied element-wise to the data using the basis functions. For a specific example, let's consider a quadratic basis function for linear regression:

$$\phi(x) = [1, x, x^2]$$

Assuming a dataset with a single feature x, the linear regression model with the quadratic basis function is given by:

$$y(\mathbf{w}, x) = w_0 + w_1 x + w_2 x^2$$

where $\mathbf{w} = [w_0, w_1, w_2]$ are the weights associated with the basis functions.

The predicted output $\hat{y}$ is obtained by applying the basis function to the input data:

$$\hat{y} = \phi(x) \cdot \mathbf{w} = w_0 + w_1 x + w_2 x^2$$

This represents a quadratic regression model where the relationship between x and $\hat{y}$ is not strictly linear but includes quadratic terms.

1.1.2 Choice of Basis Function

In general, the choice of basis function depends on the characteristics of the data and the type of relationships we want to capture. Other basis functions, such as sinusoidal or exponential functions, can also be used to model more complex patterns in the data. The key idea is to transform the input features using these basis functions to enable the model to capture non-linear relationships.

1.2 Complexity vs. Generalization Trade-off

However, with great flexibility comes the challenge of determining the complexity of the basis functions. Striking the right balance is crucial, as overly complex functions may lead to overfitting on the training data, compromising generalizability.

1.3 Power of Linear Regression

It's important to note that linear regression, when augmented with basis expansion, becomes a significantly powerful tool. The transformed data can now be modeled in a more intricate and nuanced manner, capturing non-linear relationships.

In summary, linear regression, when extended through basis expansion, transcends its apparent limitations and provides a versatile framework for modeling diverse datasets.

2 Classification

2.1 Understanding Classification: Intuition and Applications

Goal: Gaining Intuition

In the realm of classification, the primary goal is to develop an intuition for effectively separating two classes - positive and negative. This foundational concept is instrumental in various real-world applications.

Applications of Classification

Classification is a powerful tool with a multitude of applications, including:

- **Spam vs. Not Spam** Distinguishing between spam and non-spam emails is a classic use case. Training a classifier helps automate this process for efficient email filtering.
- **Fraud vs. Not Fraud** Building a classifier to differentiate between fraudulent and non-fraudulent transactions is essential for financial security systems.
- **Content Moderation** Predicting the toxicity of online posts aids in content moderation, fostering a safer online environment.
- **Object Recognition** Classifying images, such as distinguishing between Chihuahuas vs. blueberry muffins, showcases the versatility of classification.
- **Cancer vs. Not Cancer** Cancer detection involves training a classifier to differentiate between cancerous and non-cancerous cases. This extends to multi-class classification for different cancer stages.

Challenges and Dangers in Classification

While classification is a powerful tool, it comes with challenges and potential pitfalls:

- **Spurious Predictors** Instances where the model unintentionally picks up on irrelevant features. An example includes a classifier trained to detect skin lesions that mistakenly associates the presence of rulers in images with the likelihood of a lesion, overlooking the true underlying cause.
- **Discrimination Concern:** Classification models can inadvertently perpetuate biases, leading to gender or race discrimination. It is crucial to be aware of and mitigate these issues in model development.

2.2 Options for Classification Loss Functions

Option 1: 0/1 Loss

The simplest form of classification loss is the 0/1 loss, where the penalty is 0 for correct predictions and 1 for incorrect predictions.

$$0/1 \text{ Loss} = \begin{cases} 0 & \text{if } \hat{y} = y \text{ (correct prediction)} \\ 1 & \text{if } \hat{y} \neq y \text{ (incorrect prediction)} \end{cases}$$

However, this binary nature poses challenges. The 0/1 loss is not differentiable, making it unsuitable for optimization. Derivatives are essential for gradient-based optimization algorithms. The lack of differentiability implies a non-informative gradient, meaning there is no partial credit assigned for predictions that are close but not identical to the ground truth. It is analogous to a pass/fail scenario.

Option 2: Hinge Loss

To address differentiability concerns, the hinge loss is introduced. It is similar to the 0/1 loss but introduces a penalty z that depends on the degree of disagreement. The hinge loss is defined as:

$$\max(0, z)$$

If z is negative, the maximum is 0, and if z is positive, the maximum is z. This function is known as the ReLU (Rectified Linear Activation Unit) function.

Hinge loss is continuous, allowing for differentiation. This enables gradient-based optimization, a crucial aspect for training machine learning models. Only instances with a loss contribute to the overall loss. It distinguishes between correctly classified instances (loss = 0) and incorrectly classified instances (loss = z).

Option 3: Fisher's Discriminant

Fisher's discriminant is a technique associated with linear discriminant analysis. The goal is to find a projection that maximizes the distance between class means while minimizing the within-class variance.

Let $I_1 = \{n : y_n \in c_1\}$ and $I_2 = \{n : y_n \in c_2\}$ represent index sets for points in classes c_1 and c_2 respectively, where $|I_1| = N_1$ and $|I_2| = N_2$.

Characterize points through mean and variance:

$$m_1 = \frac{1}{N_1} \sum_{n \in I_1} x_n \quad m_2 = \frac{1}{N_2} \sum_{n \in I_2} x_n$$

$$S_1 = \frac{1}{N} \sum_{n \in I_1} (x_n - m_1)(x_n - m_1)^\top \quad S_2 = \frac{1}{N} \sum_{n \in I_2} (x_n - m_2)(x_n - m_2)^\top$$

The within-class scatter matrix is $S_w = S_1 + S_2$, and the between-class scatter matrix is $S_B = (m_1 - m_2)(m_1 - m_2)^\top$.

The means and variances of points projected on w are given by:

$$\mu_1 = \frac{1}{N_1} \sum_{n \in I_1} w^\top x_n = w^\top m_1 \quad \mu_2 = w^\top m_2$$

$$v_1^2 = w^\top S_1 w \quad v_2^2 = w^\top S_2 w$$

The Fisher's linear discriminant is defined as:

$$\mathcal{L}(w) = \frac{-(\mu_1 - \mu_2)^2}{v_1^2 + v_2^2}$$

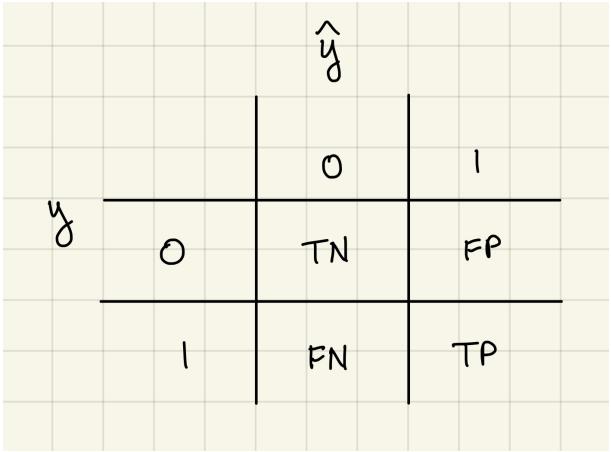
This leads to the optimal $w \propto S_w^{-1}(m_1 - m_2)$, providing the direction that maximizes the Fisher's discriminant. This ratio is maximized when between-class variance is large relative to within-class variance, resulting in a linear combination of features that effectively separates classes.

When dealing with multi-class classification problems, Fisher's discriminant is particularly valuable and it provides a principled approach to finding the optimal linear transformation for enhanced class separability.

3 Metrics to Evaluate Performance

Once we have a classifier, how do we think about performance?

3.1 Common Metrics



TN (True Negative), FP (False Positive), FN (False Negative), and TP (True Positive) are terms commonly used in binary classification to evaluate the performance of a classification model.

True Negative (TN)

Definition: Instances that are actually negative and are correctly predicted as negative by the model.

Example: In a medical diagnosis scenario, TN represents healthy patients correctly identified as healthy.

False Positive (FP)

Definition: Instances that are actually negative but are incorrectly predicted as positive by the model.

Example: In a medical diagnosis scenario, FP represents healthy patients incorrectly classified as having a disease.

False Negative (FN)

Definition: Instances that are actually positive but are incorrectly predicted as negative by the model.

Example: In a medical diagnosis scenario, FN represents patients with a disease incorrectly classified as healthy.

True Positive (TP)

Definition: Instances that are actually positive and are correctly predicted as positive by the model.

Example: In a medical diagnosis scenario, TP represents patients with a disease correctly identified as having the disease.

Furthermore, we can use these four metrics to construct additional metrics.

3.2 Precision and Recall

Precision

$$\frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It focuses on the accuracy of positive predictions made by the model. A high precision indicates that the model is making accurate positive predictions and has a low rate of false positives.

Recall (Sensitivity or True Positive Rate)

$$\frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall is the ratio of correctly predicted positive observations to the total actual positives. It focuses on the ability of the model to capture all positive instances. A high recall indicates that the model is effectively capturing most of the actual positive instances and has a low rate of false negatives.

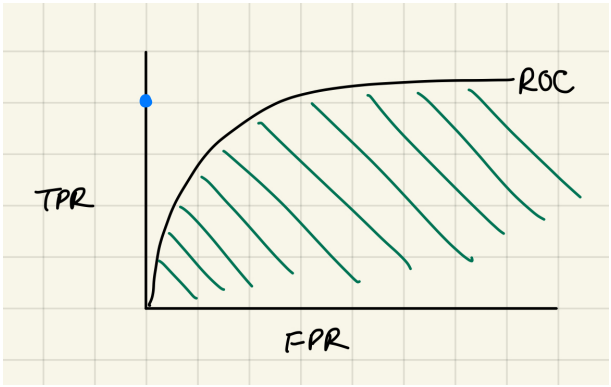
3.3 Receiver Operating Characteristic (ROC) Curve

The ROC curve is created by plotting TPR against FPR at various classification thresholds. TPR measures the proportion of actual positive instances correctly identified by the model.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

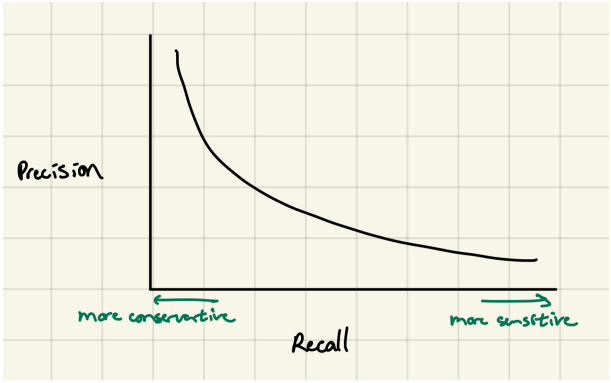
FPR measures the proportion of actual negative instances incorrectly identified as positive by the model.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$



The blue dot is where we want to be, but it is impossible so we aim to have an ROC curve that gets us as close as possible to TPR = 1 and FPR = 0.

3.4 PR Curve



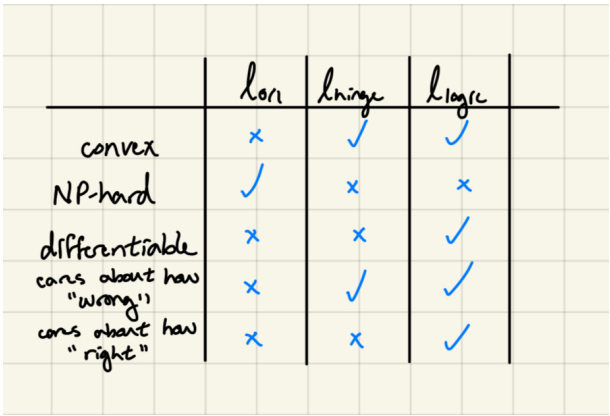
In a precision vs. recall graph, we see the trade-off between making conservative predictions and maintaining high precision in positive classifications. The further right we go, the more sensitive the model is, meaning the model is identifying a larger proportion of actual positive instances, resulting in higher recall. However, moving towards the right on the precision-recall graph often comes at the expense of precision. The further left we go, the higher the precision, but the lower the recall, meaning our predictions are more conservative.

4 Concept Check

Fill out the following table to describe the different types of classification loss functions.

	$\ell_{0/1}$	ℓ_{hinge}	ℓ_{logit}
convex			
NP-hard			
differentiable			
cares about how "wrong"			
cares about how "right"			

Solution on the following page.



Note that the l_{hinge} could also be considered differentiable if we exclude the singular point that is not differentiable. The 0/1 loss will not care how "right" or "wrong" a model prediction is, it only cares about whether or not the prediction is right or wrong. The l_{hinge} and l_{logic} graphs both give a higher penalty the more wrong an output is, so the functions "care" about how wrong an output is. Only the l_{logic} graphs "cares" about how right an output is.