Introduction to Machine Learning

Gnkgo, Informatik B. Sc. 4. Semester

Contents

1 Convexity

- If f is a differentiable convex function and $\nabla f(w) = 0$, then w is the global minimum of f.

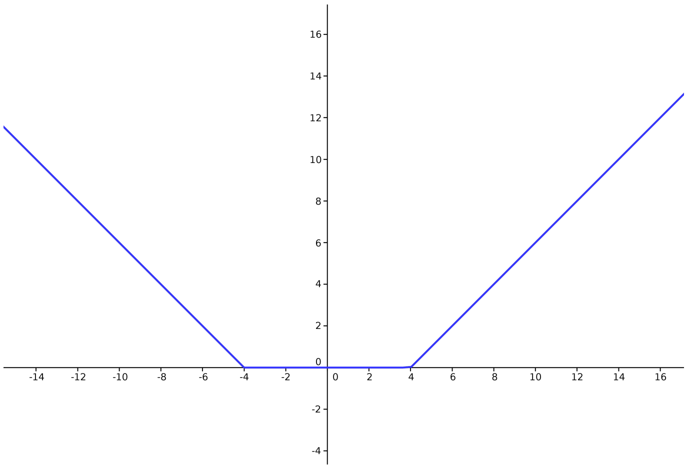


Figure 1: Convex function illustration

- Even if it is not strongly convex, it has a global minimum, just not only one.
- **Attention:** Just being differentiable and convex doesn't mean it has a stationary point: $f(w^{t+1}) < f(w^t)$.

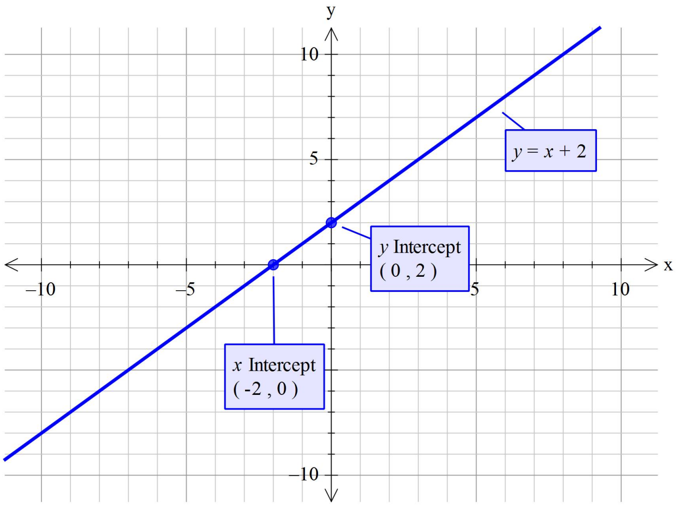


Figure 2: Convex but has no maximum, minimum, saddle point

- Only a strong convex function implies a semi-definite positive Hessian matrix.

2 Gradient Descent

Consider the gradient descent algorithm for minimizing a differentiable function f with iterates $w^{t+1} = w^t - \eta \nabla f(w^t)$. Suppose that $||\nabla f(w^t)|| > 0$. Then there always exists a step-size $\eta > 0$ such that.

Attention: This is only the case for gradient descent and not stochastic gradient descent!

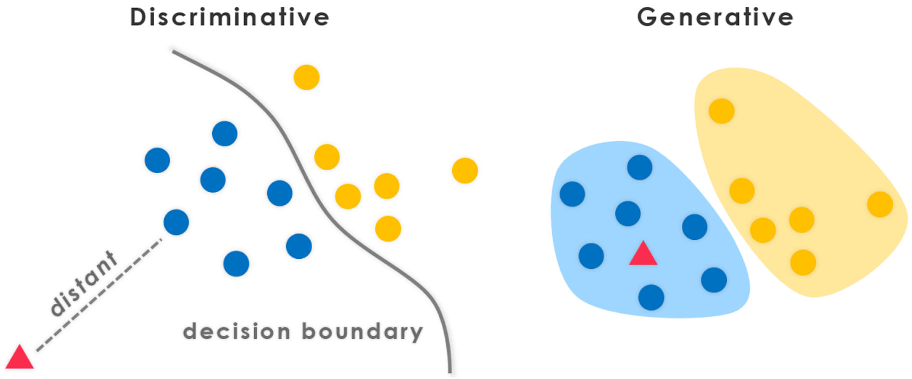


Figure 3: Difference between discriminative and generative models

3 Discriminative vs Generative Models

Description	Discriminative	Generative
What is modeled	$P(y\|\|x)$	$P(x,y)$
What is learned	Decision boundary	Probability distribution of data
Example	SVM, logistic regression	Gaussian Bayes classifier, GANS
Advantage	Cheaper, less prone to overfitting	Good at detecting outliers, generate new data

4 Gaussian Bayes Classifier (GBC)

How is $P(x,y)$ modeled?

$$P(x,y) = P(y) \cdot P(x|y)$$
$$P(Y=y) = \text{Categorical Distribution}$$
$$P(X=x|Y=y) = \text{XI}(x; M_y, \sum_y)$$

5 Convolutional Neural Network (CNN)

A convolutional neural network consists of an input layer, hidden layers, and an output layer. In a CNN, the hidden layers include one or more layers that perform convolutions.

In a CNN, the input is a tensor with shape: (number of inputs) × (input height) × (input width) × (input channels). After passing through a convolutional layer, the image becomes abstracted to a feature map, also called an activation map, with shape: (number of inputs) × (feature map height) × (feature map width) × (feature map channels).

$$\text{Parameters} = K \times K \times K \times C \times F$$

A higher threshold leads to fewer positive predictions, reducing the false positive rate for higher thresholds.

6 Ridge Regression

- Has increased bias for decreased variance.
- Closed form: $w^{\text{ridge}}(\lambda) = (X^T X + \lambda I^d)^{-1} X^T y$
- Has very low weighted values.
- Regularization tries to keep weights small.

7 Lasso Regression

- Has no closed-form solution.
- Has zero values.

8 Ordinary Least Squares

- Augmenting the set of features used for the regression will never increase the least squares loss.
- Subtracting the empirical mean from the data before performing regression on the centered samples.

9 Support Vector Machine (SVM)

- Support vectors are the closest to the boundary.
- Unconstrained soft-margin SVM is an l_2-penalized hinge loss.

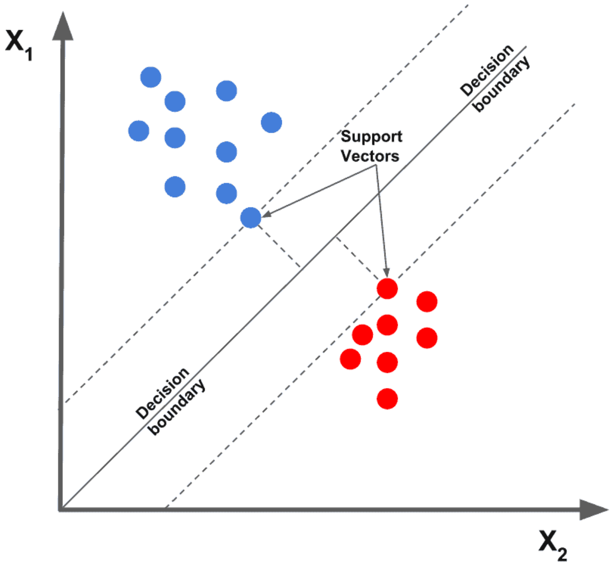


Figure 4: Support Vector Machine

10 Expectation-Maximization (EM) Algorithm

- EM algorithm converges to a local maximum/saddle point, not only with careful initialization.
- Every iteration of the EM algorithm increases the marginal likelihood of the data.
- Instead of the EM algorithm, it is possible to adapt gradient descent for learning the parameters of the GMM and its latent assignments.
- Doesn't have step size.
- Iterative optimization algorithm used to estimate the parameters of probabilistic models when some data is missing or unobserved.

11 Gaussian Mixture Model

- Probabilistic model used for representing complex data distributions.
- Works well when data is believed to be generated from a mixture of Gaussian distributions.
- Parameters of GMM:
  - Means: Represent the center of each component.
  - Covariance: Controls the shape and orientation of the component.
  - Mixing coefficients: Relative contribution of each component to the overall distribution.
- Trained using Expectation-Maximization (EM) algorithm.

12 Bootstrap

12.1 Advantage of using bootstrap parameter estimates in comparison with distribution-dependent parameter estimates

- There is no closed-form solution for bootstrap parameter estimates.
- Bootstrap sampling is a way of artificially creating more datasets. Basically, you take random samples from the dataset with replacement.
- Sampling with replacements makes it
- computationally expensive.
- Bootstrapping is possible for any ML technique, as it can be computed for any black-box predictor.
- Bootstrap estimates are not asymptotically stable.

13 Generative Adversarial Networks

D: discriminator G: neural network generator

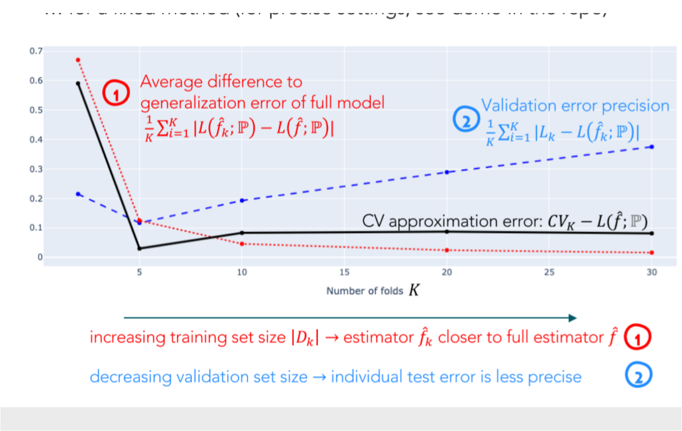
- If D and G both have enough capacity, i.e., if they can model arbitrary functions, the optimal G will be such that $G(z) \sim p_{data}$.
- The objective can be interpreted as a two-player game between G and D.
- The output of the discriminator is the probability of classifying x as being real:

$$1 - D_G(x)$$

14 Naive Bayes Classifiers

- Every pair of features being classified is independent of each other.
- Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



increasing training set size $|D_k|$ → estimator $\hat{f}_k$ closer to full estimator $\hat{f}$

decreasing validation set size → individual test error is less precise

Figure 5: Error

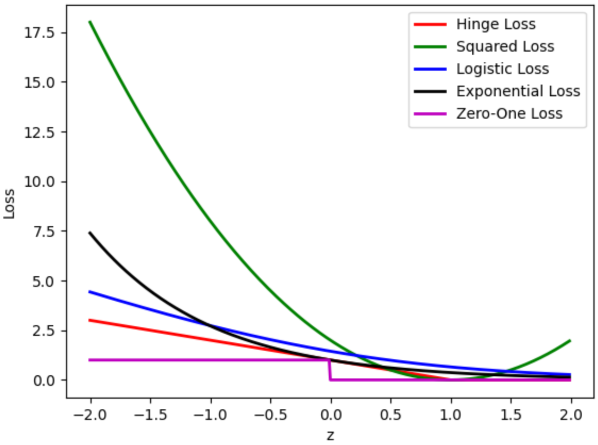


Figure 6: Error types

15 Error

- **Logistic**: Minimum is at ∞.
- **Square**:
  - Well-defined minimum, but the point is that this minimum (at 1) seems a bit random and does not make a lot of sense.
- **Exponential**:
  - Penalizes wrong labels very much and very quickly. Even one error could heavily penalize your model.
  - Has exploding derivatives for wrong results and therefore is unstable.
- **Hinge**:
  - For SVM.
  - Is convex.
  - Has a minimum.
  - Not differentiable at 1.
- **Logistic**:

  - For cross-entropy.
  - Differentiable at all points.
  - Models conditional probability $p(y|w,x)$.
  - The logistic loss doesn't necessarily maximize the margin between classes since it takes into account all the samples in both classes.

- **Linear**:
  - Too sensitive to outliers and returns garbage when there is an imbalance in data.

- **0-1-Loss**:
  - Derivative is always 0, doesn't make sense to optimize that.

- **Cross-Entropy Loss in Classification**:
  - Cross-entropy loss is a crucial component in training classification models.
  - It quantifies the dissimilarity between predicted class probabilities and actual class labels.
  - For each data point, the cross-entropy loss is computed by taking the negative logarithm of the predicted probability assigned to the true class:

$$L_i = -\sum_{k=1}^{K} y_{ik} \cdot \log(p_{ik})$$

  - This loss function not only measures the correctness of the model's predictions but also encourages the model to be confident and accurate in its class probability assignments.
  - The overall objective during training is to minimize the mean cross-entropy loss across the dataset:

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i$$

16 Asymmetric 0-1 Loss with Abstention

We shall define a new loss named 0-1 loss with abstention with an *extended action space*:

$$f(x) \in \{-1, +1, r\}$$

where r indicates **abstaining from a prediction**. This method is sometimes called **selective classification**. We also introduce a cost $c \in [0, 0.5]$ for abstaining. The loss becomes:

$$l(f(x), y) = \mathbf{1}_{f(x)\neq y}\mathbf{1}_{f(x)\neq r} + c\mathbf{1}_{f(x)=r}$$

We should abstain if:

$$c < \min\{p(x), 1 - p(x)\}$$

17 Classification

The margin of a decision hyperplane to be the (smallest) distance between the hyperplane and the data points. The margin of the hyperplane $\hat{w}$ is defined as $\frac{1}{||\hat{w}||}$.

18 Quiz

18.1 K-means clustering

- Seeks cluster centers and assignments to minimize the within-cluster sum of squares.
- Appropriate if the underlying clusters are separable, spherical, and approximately of the same size.
- K-means clustering can be kernelized.

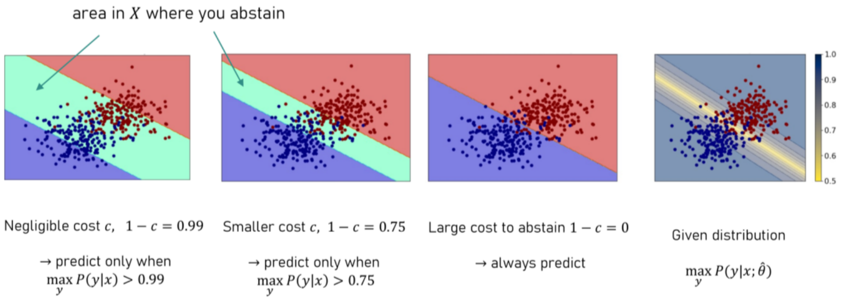


Figure 7: Asymmetric 0-1 Loss with Abstention

18.2 Find k

- By using a heuristic like the elbow method that identifies the diminishing returns from increasing k.
- By using an information criterion that regularizes the solution to favor simpler models with lower k.

18.3 Lloyd's algorithm

- It cannot cycle; i.e., it does never return to a particular solution after having previously changed to a different solution.
- Using specialized initialization schemes (e.g., k-means++) can improve the quality of solutions found by the algorithm and reduce its runtime.
- Center of clusters should be at the center of gravity.
- So after choosing centers and clustering, move centers to new centers.
- Repeat until done.
- Converges, local or global minimum.

18.4 PCA

- PCA can be kernelized.
- Unsupervised learning algorithm.
- It is orthogonal to all other principal components found by PCA.
- If we use the Gaussian kernel for kernel PCA, we implicitly perform PCA on an infinite-dimensional feature space.
- Gaussian kernel has infinite dimensions.
- Autoencoders and PCA are the same thing if we choose the activation function $\varphi(\cdot)$.
- For every arbitrary finite dataset with two classes and distinct points, there exists a feature map ϕ, such that the dataset becomes linearly separable.
  - As long as it is finite with two datasets A, B to separate, one can literally define a feature map:

$$\phi(x) = \begin{cases} 1 & \text{if } x \in A \\ -1 & \text{otherwise} \end{cases}$$

18.5 PCA first principal component

- Captures the maximum amount of variance in the data among all possible linear combinations of the original features.
- Represents the direction in the data space along which the data exhibits the highest variability or spread.
- Orthogonal to all other subsequent principal components, meaning it is uncorrelated with them. This orthogonality property allows PCA to create uncorrelated features.
- The first principal component is given by the eigenvector of the data covariance matrix with the largest eigenvalue.