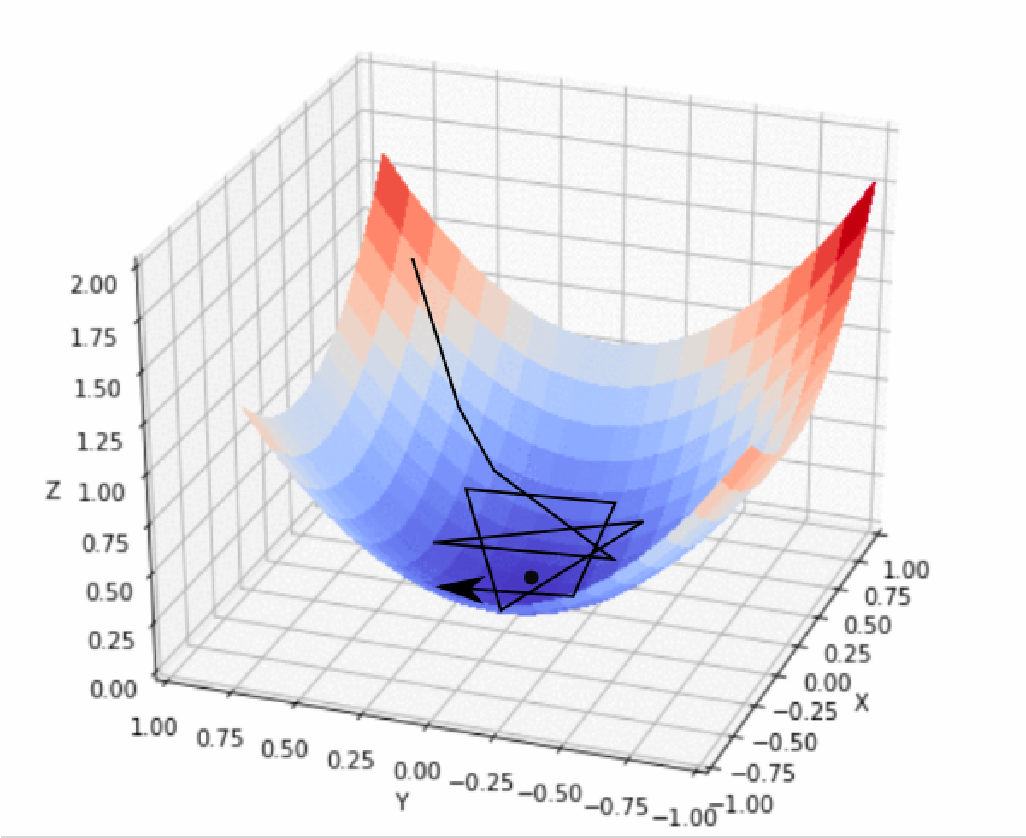
Introduction to Machine Learning

Lecture 7
Linear Models III
Classification

Goker Erdogan
26 – 30 November 2018
Pontificia Universidad Javeriana

Gradient descent

- A simple technique to solve optimization problems
  - Idea: Look around and move in the direction of steepest descent
  - Direction of steepest descent = Negative gradient



[1]

Gradient descent

Problem: $$\min_{w} \quad E(w)$$

GRADIENT DESCENT ALGORITHM

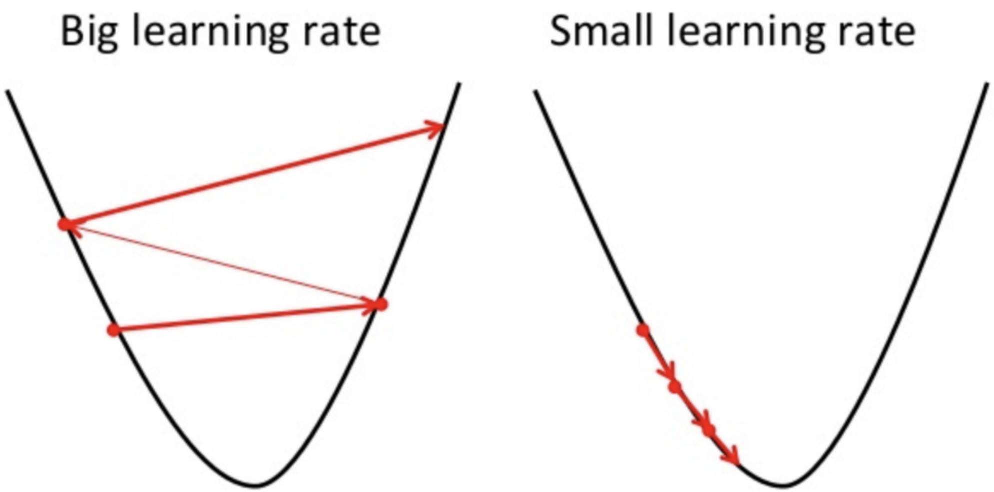
- Given a training set of N samples, {x, t}. Step size γ
- Initialize w_0 randomly
- Repeat for e=1, 2, 3, …

$$w_{e+1} = w_e - \gamma \nabla E(w)$$

- (Almost) general purpose
  - Only needs the gradient of error function
- Works well in practice

Gradient descent

- Need to pick γ carefully



[2]

- There are many variants of gradient descent
  - Momentum
  - Stochastic/minibatch GD
  - RMSprop/AdaGrad/Adam

Solving logistic regression with gradient descent

- We need $\nabla_w E(w)$

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

$$y_n = \sigma(w_0 + \sum_d w_d \phi_d(X_{nd}))$$

- Note

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$$

- Then

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n$$

Two perspectives on ML

Function approximation perspective

1) Assume a parametric relationship $f_\theta: x \rightarrow y$
   - θ are the parameters
2) Define a function that measures how good/bad some θ is

   Error function: $E(\theta)$

3) Solve

   $\min_\theta \quad E(\theta) + \lambda R(\theta)$

Regularization

The probabilistic perspective - basics

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

[5]

- Bayes' theorem

Likelihood

Prior

$$p(\theta|D = (x, y)) = \frac{p(D = (x, y)|\theta)p(\theta)}{p(D)}$$

Posterior

The probabilistic perspective

Probabilistic (maximum-likelihood) perspective

1) Assume a parametric distribution $y \sim f_\theta(x)$
   - θ are the parameters
2) Write the likelihood of θ given data x, y

   Likelihood function: $p(Y=y|\theta, X=x)$

3) Maximize likelihood

   $\max_\theta \quad p(Y=y|\theta, X=x)$

Maximum-a-posteriori estimation

– Assume a prior $p(\theta)$ and maximize posterior

   $\max_\theta \quad p(Y=y|\theta, X=x)\, p(\theta)$

Regularization

Logistic regression revisited

- Let's look at binary classification

1) Assume a Bernoulli distribution for t given x

$$t \sim \mathrm{Be}(\sigma(w^T x))$$

σ : Sigmoid function

$$p(t = 1|w, x) = \sigma(w^T x)$$

- If y ~ Be(p), then y=1 with probability p
- For example, a fair coin would have Be(0.5) distribution

2) Write the likelihood function p(y|w, x) for all data

- Given N training samples $\{x, t\}_{n=1...N}$

- $\{x_1, t_1 = 1\}$ $\longrightarrow$ $\sigma(w^T x_1)$
- $\{x_2, t_2 = 0\}$ $\longrightarrow$ $1 - \sigma(w^T x_2)$
- … …
- $\{x_N. t_N = 0\}$ $\longrightarrow$ $1 - \sigma(w^T x_N)$

Logistic regression revisited

- In other words, we want
  - $\sigma(w^T x)$ if t=1
  - $1 - \sigma(w^T x)$ if t=0

- A clever way to write that

$$p(t|w,x) = \sigma(w^T x)^t (1 - \sigma(w^T x))^{(1-t)}$$

- Then the likelihood function

$$p(\mathbf{t}|\mathbf{w}, X) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{(1-t_n)}$$

  - where

$$y_n = \sigma(w^T X_n)$$

Logistic regression revisited

3) Finally, maximize likelihood function

$$\max_{\mathbf{w}} \quad p(\mathbf{t}|\mathbf{w}, X) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{(1-t_n)}$$

- Take log and multiply by -1

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

- Voila!

Minimize cross entropy loss = Maximize likelihood (under Bernoulli model)

Linear regression revisited

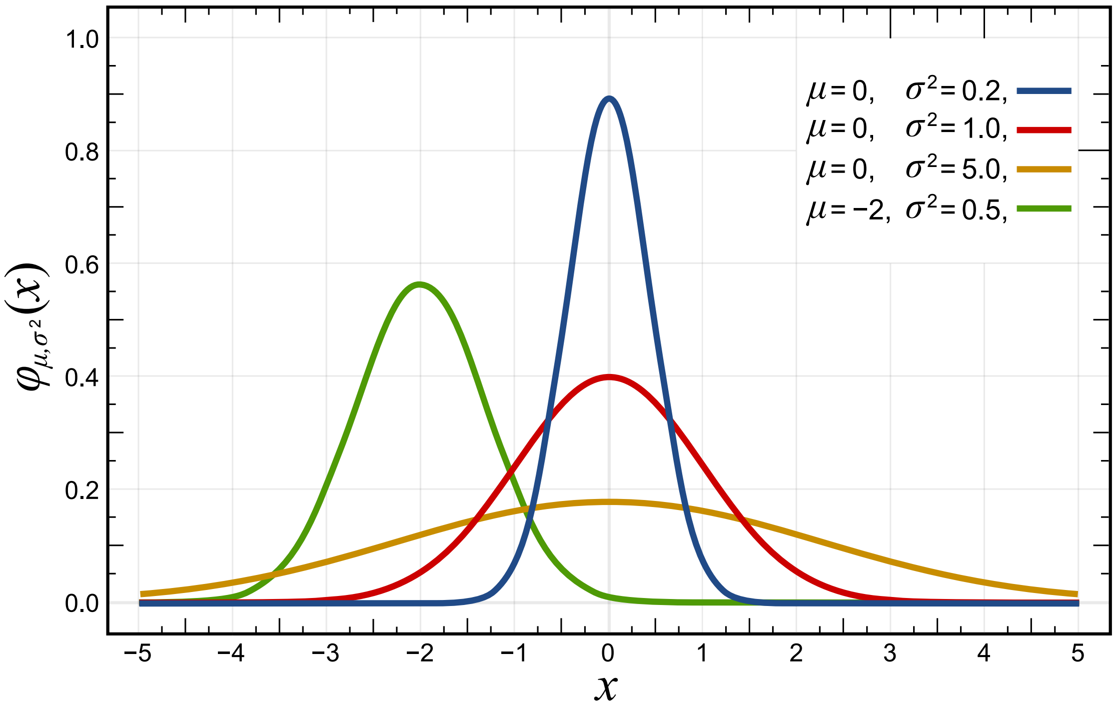
1) Assume a Gaussian distribution for t given x

$$t \sim \mathcal{N}(w^T x, \sigma^2)$$

mean

σ^2: Variance

$$p(t|w,x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-w^T x)^2}{2\sigma^2}}$$



[4]

Linear regression revisited

- Write likelihood function

$$p(\mathbf{t}|\mathbf{w}, X) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^T x, \sigma^2)$$

- Take log and multiply by -1 (and get rid of terms that do not depend on w)

$$E(w) = -\ln p(\mathbf{t}|\mathbf{w}, X) = \sum_{n=1}^{N} (t_n - w^T x)^2$$

Minimize sum-of-squares = Maximize likelihood (under Gaussian model)

Regularization from a probabilistic perspective

Maximum-a-posteriori estimation

– Assume a prior p(θ) and maximize posterior

$\max_\theta \quad p(Y=y|\theta, X=x)\, p(\theta)$

Regularization

– For linear regression, assume a Gaussian prior on w

$$\mathbf{w} \sim \mathcal{N}(0, \alpha^2 \mathbf{I})$$

– Then maximum-a-posteriori estimation problem is ridge regression (l_2)

$$\begin{aligned} E(\mathbf{w}) &= -\ln p(\mathbf{t}|\mathbf{w}, X) - \ln p(\mathbf{w}) \\ &= \sum_{n=1}^{N} (t_n - w^T x)^2 + \lambda \sum_d w_d^2 \end{aligned}$$

Why take all this trouble?

- Uncertainty is fundamental in ML
  - Noise on measurement
  - Finite size of datasets
- Probability theory is the calculus of uncertainty
  - Extension of logic to situations involving uncertainty
- Allows us to measure confidence in predictions
- Offers a principled way to formulate problems
  - Learning = Probabilistic inference
  - Allows us to adapt/generalize models
  - e.g., K-means → Gaussian mixture models
- Most ML techniques are probabilistic

Summary

- Gradient descent
  - Solving logistic regression
- Two perspectives on ML
- Probabilistic perspective
  - Maximum-likelihood and maximum-a-posteriori
  - Logistic regression from a probabilistic perspective
  - Linear regression from a probabilistic perspective

- Exercises
  - Derive the gradient of error function for binary logistic regression
  - Show that negative log likelihood for linear regression is the sum-of-squares error function

References

[1] https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

[2] https://towardsdatascience.com/gradient-descent-in-a-nutshell-eaf8c18212f0

[3] Bishop. Pattern Recognition and Machine Learning. Chapter 4.

[4] https://en.wikipedia.org/wiki/Normal_distribution

[5] Ghahramani. http://mlg.eng.cam.ac.uk/zoubin/talks/mit12csail.pdf