The general paradigm of supervised learning

- The goal of a large family of machine learning is to minimize the prediction errors of the model
  - Ideally we want to predict the true errors, err model when it is used in real-life scenarios
  - That is hard to do, so the common practice is to minimize the errors in a t
  - To do that we ne n, which is a metric that measures the errors in the p l.
    - Cross-entropy Loss, Squared E
- In other cases it is more natural to think of th is to optimize an *objective function*, e.g., Maximum Likelihood
- Whether to call is a loss function or objective function, there is no difference in how they are optimized

Commonly used loss and objective functions in NLP

- Naïve Bayes: m https://eduassistpro.github.io/t of labeled samples

Assignment Project Exam Help

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\boldsymbol{x}^{1:N}, y^{1:N}; \boldsymbol{\theta})$$

Add WeChat edu_assist_pro

Assignment Project

- Logistic Regression: The weights are estim **Maximum Condition**

https://eduassistpro.github.io/

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\text{argmax}} \log$$

Add WeChat edu_assist_pro

- SVM: The weights are estimated by minimizing marginal loss

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{i=1}^{N} \left(1 - \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})\right)_+$$

Note: Letters in bold indicates vector: $\boldsymbol{\theta}$, $\boldsymbol{x}$, $\boldsymbol{f}$. Alternative notations: $\vec{\theta}$, $\vec{x}$, $\vec{f}$

Naïve Bayes Objective

- Naïve Bayes: Maximize the joint probability of a training set of labeled samples, in a process called **ikelihood Estimation**

$$= \operatorname*{argmax}_{\theta} \prod_{i=1}^{N} P(\boldsymbol{x}^i, y^i; \boldsymbol{\theta})$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log P(\boldsymbol{x}^i, y^i; \boldsymbol{\theta})$$

Logistic Regression Objective

- Logistic Regression: The weights are estimated by **Maximum Conditional Likelihood**

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(y^{1:N} | \boldsymbol{x}^{1:N}, \boldsymbol{\theta})$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y \in \mathcal{Y}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y) \right) \right)$$

or by minimizing the **logistic loss**

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} - \sum_{i=1}^{N} \left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y \in \mathcal{Y}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y) \right) \right)$$

Support Vector Machine Objective

- SVM: The weights are estimated by minimiz

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}}$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{i=1}^{N} (\max_{y \in \mathcal{Y}} (\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, \quad) + \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}))_+$$

These look rather daunting, don't they?

How do we minimize a function?

In order to minimize a function, we need to be able to co
derivative, or rate of change of the function.

Let's start with a much simpler function nd its derivative is:

$$\frac{d}{dx} f(x) = \frac{d}{dx}(x^2$$

"The derivative of the function $f(x)$ with respect to (w.r.t.) x"
This looks like magic, but it's really just calculus.

How do we find the minimum of a function with the derivative?

- The derivative of a function can be interprete
  certain point of the function.
- At the point that is the minimum (or maximum) of the function, the s
  We can find the
  derivative to zero:
  $2x = 0, x = 0$
- For this particular function, there is a closed form solution. Most models in NLP don't have a closed form solution, but some do, e.g., Naïve Bayes.

Plot the function

Finding the minimum iteratively

For functions that don't, we can find the minimum iteratively.

from the input x so that the value of the function will decrease.

Suppose we start at the point where $x = \quad 1$, and set the fraction $\eta \triangleq 0.1$, and $\Delta x = \eta \frac{d}{dx} f(x)$. So:

$x = x$

$f(x) = (-0.8) \quad + 1 = 1.64$

$x = x - \Delta x = -0.8 - 0.1 \quad 64$

$f(x) = (-0.64)^2 + 1 = 1.4096$

$x = x - \Delta x = -0.64 - 0.1 \times (-1.28) = -0.512$

$f(x) = (-0.512)^2 + 1 = 1.262144$

As x approaches 0, $f(x)$ reaches the minimum, which is 1.

Finding the minimum iteratively

Finding the minimum iteratively

What if we try to learn fast using a larger learning rate?

Let's still start at x the learning rate $\eta \triangleq 1$ instead and see what hap

$x = x - \Delta x = -1 - 1$

$f(x) = (-1)^2 + 1 = 2$

$x = x - \Delta x = -1 -$

$f(x) = (1)^2 + 1 = 2$

So the x will just swing back and forth without ever reaching the minimum.

Setting the right learning rate is thus very important. If set improperly, we'll never reach the minimum, or at least take much longer than necessary.

Trying to learn fast with a larger learning rate

Trying to learn fast with a larger learning rate

Derivative Rules

Common Derivatives

$$\frac{d}{dx}(C) = 0 \qquad e.g., \frac{d}{dx}(91) = 0$$

$$\frac{d}{dx}(x) = 1 \qquad e.g., \frac{d}{dx}(x) = 1$$

$$\frac{d}{dx}(x^n) = nx^{n-1} \qquad e.g., \frac{d}{dx}(x^4) = 4x^3$$

$$\frac{d}{dx}(a^x) = a^x \ln(a) \qquad e.g., \frac{d}{dx}(5^x) = 5^x \ln(5)$$

$$\frac{d}{dx}(e^x) = e^x$$

Note: ln: "Natural logarithm", logarithm to base of the mathematic constant e, where $e = 2.71882\cdots$

Derivative rules

More common derivat

$$\frac{d}{dx}(\ln(x)) = \frac{1}{x}, x > 0$$

$$\frac{d}{dx}(\ln(|x|)) = \frac{1}{x}, x \neq$$

$$\frac{d}{dx}(\sin(x)) = \cos($$

$$\frac{d}{dx}(\cos(x)) = -\sin$$

$$\frac{d}{dx}(\tan(x)) = \sec^2(x)$$

Note: When $x <= 0$, $\ln(x)$ is unspecified. That is, you can't raise the constant e to any value to get a zero or a negative number.

Derivatives of functions

"The derivative of the function with respect to x"

$$\frac{d}{dx}(cf(x)) = c\frac{d}{dx}f(x)$$

$$\frac{d}{dx}(f(x) \pm g(x)) = \frac{d}{dx}f(x) \pm \frac{d}{dx}g(x)$$

$$\frac{d}{dx}(f(x)g(x)) = f(x)\frac{d}{dx}g(x) + g(x)\frac{d}{dx}f(x) \quad \text{(Product rule)}$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{g(x)\frac{d}{dx}f(x) - f(x)\frac{d}{dx}g(x)}{g^2} \quad \text{(Quotient rule)}$$

$$\frac{d}{dx}f(g(x)) = \frac{d}{dg(x)}f(g(x))\frac{d}{dx}g(x) \quad \text{(Chain rule)}$$

Breaking down the derivative of complex functions

Using these fundamental derivative rules, and particularly the chain rule, you can break down more complicated functi

$$\frac{d}{dx}e^{f(x)} = e^{f(x)}\frac{d}{dx}$$

$$\frac{d}{dx}\ln(f(x)) = \frac{1}{f(x)}\frac{d}{dx}$$

Partial Derivatives

- We don't normally deal with single variable fu
  A typical NLP model (function) has tens of tho
  millions of variables (features). So we need to compute *partial derivatives*.
- Fortunatel imple.
  You just need to hold all other variables con
  as constant), and take the derivative wit
  variable. $\frac{\partial}{\partial x} f(x, y)$, $\frac{\partial}{\partial y} f(x, y)$

More on partial derivatives

$$f(x,y) = x^2 +$$

$$\frac{\partial}{\partial y} f(x,y)$$

More on partial derivatives

$$f(x,y) = \min(x,y) = \begin{cases} x & \text{if} \\ y & \text{if} \end{cases}$$

$$\frac{\partial}{\partial x} f(x,y) = \begin{cases} 1, & \text{if} \quad x < y \\ 0, & \text{if} \quad x > y \end{cases}$$

$$\frac{\partial}{\partial y} f(x,y) = \begin{cases} 0, & \text{if} \quad x < y \\ 1, & \text{if} \quad x > y \end{cases}$$

The function is not differentiable when $x = y$

Plot multi-variable functions

Gradient

The gradient of a function ∇f is the set of pa
function

$$\nabla f = \begin{bmatrix} \frac{\partial}{\partial} \\ \vdots \\ \frac{\partial}{\partial} \end{bmatrix}$$

Properties of Logarithms

$$\log(xy) = \log(x) + \log(y) \qquad \ln(e^x) = x$$

$$\log(\frac{x}{y}) = \log(x) - \log(y) \qquad \ln(x) \qquad > 0$$

$$\log(x^y) = y\log(x)$$

$$\log(\prod_i x_i$$

- It is common practice to map probabilities to logarithmic space to avoid *underflow* (when a value gets too close to zero for the computer to represent it).
  $\ln(0.0001) = -9.2103403\cdots$
- You can map the log values back to probabilities using the exponent. $e^{-9.2103403} = 0.0001$

Convexity of functions

- Intuitively, a convex (conclave) function is a continuous function in which there is a single minimum (m
- A mathematical definition: A convex functi function who domain does n the ends of the interval.
- How to decide a function is convex. If cond derivative in $[a, b]$, then a necessary a ion for it to be convex on that interval is that the second derivative $f''(x) \geq 0$ for all x in $[a, b]$.

Example convex functions

Example non-convex functions