(Stochastic) Gradient Descent

THE AUSTRALIAN NATIONAL UNIVERSITY

Empirical Risk Functional $R_{\mathrm{emp}}[f] = \frac{1}{m} \sum_{i=1}^{m} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$

Idea 1

Minimize $R_{\mathrm{emp}}[f]$ by performing gradient descent. This leads to

$$f \to f - \frac{\Lambda}{m} \sum_{i=1}^{m} \partial_f c(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$$

Problem

This may be expensive. If the observations are similar, this is very wasteful.

Idea 2

Minimize $R_{\mathrm{emp}}[f]$ by performing stochastic gradient descent over the individual terms under the sum.

Stochastic Gradient $f \to f - \Lambda \partial_f c(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$

Linear Model $\mathbf{w} \to \mathbf{w} - \Lambda \mathbf{x}_i c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$

Perceptron Algorithm for Squared Loss

argument: Training sample, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathcal{X}$, $\{y_1, \ldots, y_m\} \subset \{\pm 1\}$, η

returns: Weight vector $\mathbf{w}$ and threshold b.

function Perceptron(X, Y, η)

  initialize $\mathbf{w}, b = 0$

  repeat

    for all i from $i = 1, \ldots, m$

      Compute $f(\mathbf{x}_i) = \left(\left\langle \sum_{l=1}^{i} \alpha_l \Phi(x_l), \Phi(\mathbf{x}_i) \right\rangle + b\right)$

      Update $\mathbf{w}, b$ according to $\mathbf{w}' = \mathbf{w} + \eta \alpha_i \Phi(\mathbf{x}_i)$ and $b' = b + \eta \alpha_i$

      where $\alpha_i = y_i - f(\mathbf{x}_i)$

    endfor

  until for all $1 \leq i \leq m$ we have $g(\mathbf{x}_i) = y_i$

  return $f : \mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b$

end

Perceptron Algorithm for Huber's Loss

THE AUSTRALIAN NATIONAL UNIVERSITY

argument: Training sample, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathcal{X}$, $\{y_1, \ldots, y_m\} \subset \{\pm 1\}$, η

returns: Weight vector $\mathbf{w}$ and threshold b.

function Perceptron(X, Y, η)

initialize $\mathbf{w}, b = 0$

repeat

for all i from $i = 1, \ldots, m$

Compute $f(\mathbf{x}_i) = \left(\left\langle \sum_{l=1}^{i} \alpha_l \Phi(x_l), \Phi(\mathbf{x}_i) \right\rangle + b\right)$

Update $\mathbf{w}, b$ according to $\mathbf{w}' = \mathbf{w} + \eta\alpha_i\Phi(\mathbf{x}_i)$ and $b' = b + \eta\alpha_i$

where $\alpha_i = \begin{cases} \frac{1}{\sigma}(y_i - f(\mathbf{x}_i)) & \text{for } |y_i - f(\mathbf{x}_i)| \leq \sigma \\ \operatorname{sgn}(y_i - f(\mathbf{x}_i)) & \text{otherwise} \end{cases}$

endfor

until for all $1 \leq i \leq m$ we have $g(\mathbf{x}_i) = y_i$

return $f : \mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b$

end

Learning Rate

THE AUSTRALIAN NATIONAL UNIVERSITY

Classification

For classification, the absolute value of f does not matter. So we need not adjust the learning rate.

Regression

The absolute value of f is crucial, so we have to get η right.

- Large η: we get quick initial convergenceto the target but large fluctuationsremain (stochastic gradient can be very noisy).
- Small η: slow initial convergenceto the target but we have a much better quality estimate in the later stages.

Trick

Make η a variable of the time. One can show that $\eta(t) = O(t^{-1})$ is optimal in many cases. This yields quick initial convergence and low fluctuations later.

Warning

If f is fluctuating, choosing η too small will not be useful.

Maximum Likelihood and Noise Models

THE AUSTRALIAN NATIONAL UNIVERSITY

Basic Idea

We assume that the observations y_i are derived from $f(\mathbf{x}_i)$ by adding noise, i.e. $y_i = f(\mathbf{x}_i) + \xi_i$ where ξ_i is a random variable with density $p(\xi_i)$.

This also means that once we know the type of noise we are dealing with, we may compute conditional densities $p(y|\mathbf{x})$ under the model assumptions.

Likelihood $p(Y|f, X) = p((y_1 - f(\mathbf{x}_1)), \ldots, (y_m, f(\mathbf{x}_m)))$

We make the assumption of iid data (to keep the equations simple). This leads to the likelihood

$$\mathcal{L} = \prod_{i=1}^{m} p(y_i - f(\mathbf{x}_i))$$

Caveat

The estimates we obtain are only as good as our initial assumptions regarding the type of function expansion and noise. This means that we may not take $p(Y|X)$ at book value.

Log-Likelihood and Loss Function

THE AUSTRALIAN NATIONAL UNIVERSITY

Idea

Log likelihhood and loss function look suspiciously similar, maybe we can find a link . . . . For simplicity we assume that the that is generated iid.

Comparison

$$
\begin{aligned}
-\mathfrak{L}[f] &= \sum_{i=1}^{m} \log p(y_i - f(\mathbf{x}_i)) \\
R_{\mathrm{emp}}[f] &= \frac{1}{m} \sum_{i=1}^{m} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i))
\end{aligned}
$$

Idea

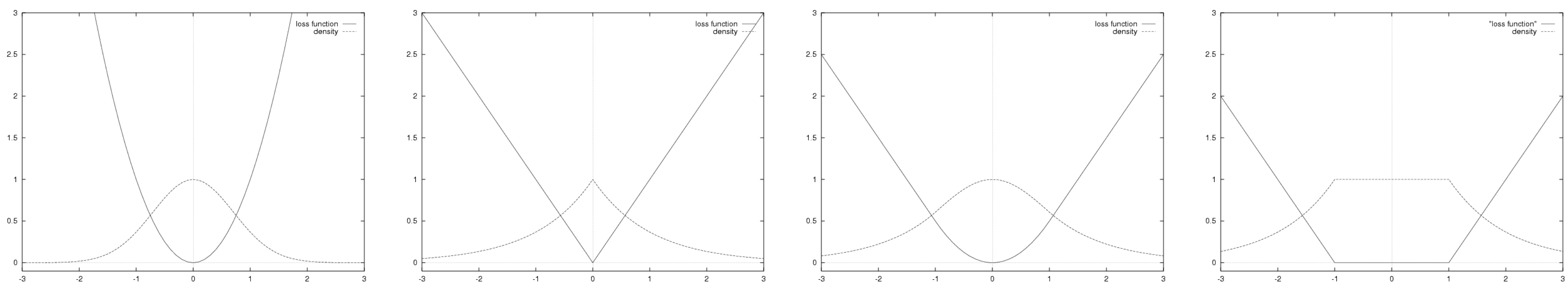
The two terms differ only by a scaling constant which is irrelevant for minimization purposes. So match up the terms.

$$
\begin{aligned}
c(\mathbf{x}, y, f(\mathbf{x})) &\equiv -\log p(y_i - f(\mathbf{x}_i)) \\
p(y_i | f(\mathbf{x}_i)) &\equiv \exp(-c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)))
\end{aligned}
$$

Density and Loss

THE AUSTRALIAN NATIONAL UNIVERSITY

	loss function $\tilde{c}(\xi)$	density model $p(\xi)$
ε–insensitive	$\lvert\xi\rvert_\varepsilon$	$\frac{1}{2(1+\varepsilon)}\exp(-\lvert\xi\rvert_\varepsilon)$
Laplacian	$\lvert\xi\rvert$	$\frac{1}{2}\exp(-\lvert\xi\rvert)$
Gaussian	$\frac{1}{2}\xi^2$	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{\xi^2}{2})$
Huber's robust loss	$\begin{cases} \frac{1}{2\sigma}(\xi)^2 & \text{if } \lvert\xi\rvert \leq \sigma \\ \lvert\xi\rvert - \frac{\sigma}{2} & \text{otherwise} \end{cases}$	$\propto \begin{cases} \exp(-\frac{\xi^2}{2\sigma}) & \text{if } \lvert\xi\rvert \leq \sigma \\ \exp(\frac{\sigma}{2} - \lvert\xi\rvert) & \text{otherwise} \end{cases}$



A Worked-Through Example, Part I

Function Expansion

We use a linear model (as in the previous lecture) $f_1, \ldots, f_n$ such that

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i f_i(\mathbf{x})$$

Additive Noise

Assume Gaussian noise ξ which corrupts the measurements such that we observe y rather than $f(\mathbf{x})$, i.e. $y = f(\mathbf{x}) + \xi$. We write $\xi \sim \mathcal{N}(0, \sigma)$ in order to state that

$$p(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\xi^2}.$$

Density Model

From above we know that $p(y|\mathbf{x}, \alpha, \sigma)$ is given by

$$p(y|\mathbf{x}, \alpha, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - f(\mathbf{x}))^2\right)$$

A Worked-Through Example, Part II

THE AUSTRALIAN NATIONAL UNIVERSITY

Likelihood

Under the assumption of iid data, the likelihood of observing $Y = \{y_1, \ldots, y_m\}$, given $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ can be found as

$$p(Y|X, \alpha, \sigma) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \alpha, \sigma)$$

Log Likelihood

$$\begin{aligned}
\mathfrak{L} &= \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \alpha, \sigma) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - f(\mathbf{x}_i))^2\right) \\
&= -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - f(\mathbf{x}_i))^2
\end{aligned}$$

A Worked-Through Example, Part III

THE AUSTRALIAN NATIONAL UNIVERSITY

Optimality Criterion

We need a maximum with respect to the parameters α, σ. The conditions $\partial_\alpha \mathcal{L} = 0$ and $\partial_\sigma \mathcal{L} = 0$ are necessary for this purpose.

Optimality in α $\partial_\alpha \mathcal{L} = \partial_\alpha \frac{1}{2\sigma^2} \|\mathbf{y} - F\alpha\|^2 = \frac{1}{\sigma^2}(F^\top F\alpha - F^\top \mathbf{y}) = 0$

Here we defined (as before) $F_{ij} = f_j(\mathbf{x}_i)$. It leads to the standard least mean squares solution $\alpha = (F^\top F)^{-1} F^\top \mathbf{y}$.

Optimality in σ

$$\partial_\sigma \mathcal{L} = \frac{m}{\sigma} - \frac{1}{\sigma^2} \sum_{i=1}^{m} (y_i - f(\mathbf{x}_i))^2 = 0$$

Likewise this leads to $\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(\mathbf{x}_i))^2$ which is *empirical* variance given by the model on the training set.

When Things go wrong with ML

THE AUSTRALIAN NATIONAL UNIVERSITY

No fine-grained prior knowledge

All functions we optimize over are treated as equally likely.

Not possible to check assumptions

- Our ML model works if the assumptions are correct. However, it breaks if they are not all satisfied. And it is hard to test them.
- Difficult to integrate alternative estimates.
- Confidence bounds for estimates.

High dimensional estimates break

- Overly confident estimates
- Overfitting
- Likelihood diverges: assume $y_i = f(\mathbf{x}_i)$. In this case we would estimate $\sigma = 0$ as the empirical variance. This in turn leads to $\mathcal{L} \to \infty$.

Regularization

Problem

The space of the solutions for f is too large if we admit all possible solutions in, say, the span of $f_1, \ldots, f_n$. Moreover we want to **rank** the solutions.

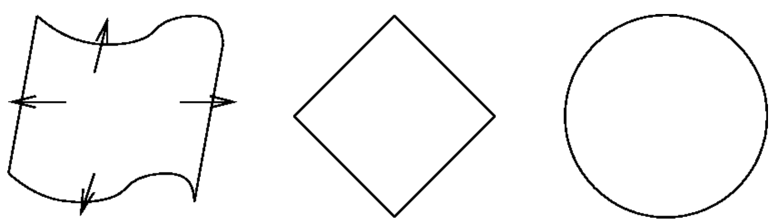
Idea

Restrict the possible solutions to the set $\Omega[f] \leq c$ where $\Omega[f]$ is some convex function

$$\Omega[f] = \sum_{i=1}^{n} |\alpha_i| \; (\ell_1 \text{ Regularization})$$

$$\Omega[f] = \frac{1}{2}\sum_{i=1}^{n} \alpha_i^2 \; (\ell_2 \text{ Regularization})$$

$$\Omega[f] = \frac{1}{2}\alpha^\top M \alpha \text{ here } M \text{ is a positive semidefinite matrix}$$



Regularized Risk Functional

THE AUSTRALIAN NATIONAL UNIVERSITY

Problem

Restricting f to the subset $\Omega[f] \leq c$ will solve the problem but the optimization problems are sometimes rather difficult to solve.

Idea

Trade off the size of $\Omega[f]$ with respect to $R_{\text{emp}}[f]$ and minimize the sum of these two terms.

Definition

For some $\lambda > 0$, also referred to as the regularization constant, the regularized risk functional is given by

$$R_{\text{reg}}[f] = R_{\text{emp}} + \lambda\Omega[f] = \frac{1}{m}\sum_{i=1}^{m} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \lambda\Omega[f]$$

This is the central quantity in most learning settings. Note that $R_{\text{reg}}[f]$ is convex, provided $R_{\text{emp}}[f]$ and $\Omega[f]$ are.

Example: Adding to the Diagonal

Quadratic Loss $c(\mathbf{x}, y, f(\mathbf{x})) = \frac{1}{2}(y - f(\mathbf{x}))^2$

Linear Model $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i f_i(\mathbf{x})$

ℓ_2 **Regularizer** $\Omega[f] = \sum_{i=1}^{n} \alpha_i^2$

Regularized Risk Functional

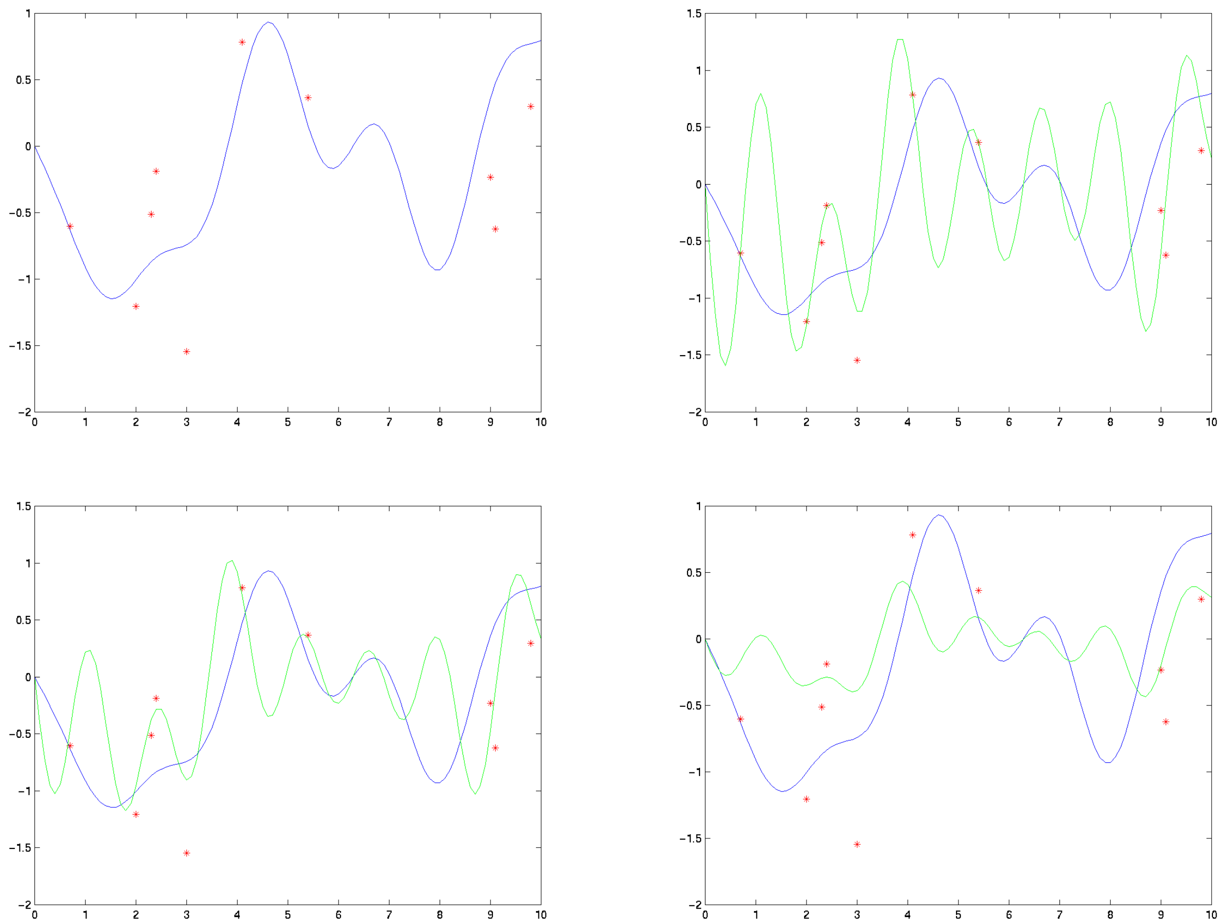
$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(y_i - f(\mathbf{x}_i))^2 + \frac{\lambda}{2} \sum_{i=1}^{n} \alpha_i^2 = \frac{1}{2m} \|\mathbf{y} - F\alpha\|^2 + \frac{\lambda}{2} \|\alpha\|^2$$

Optimality Conditions

$\partial_\alpha R_{\text{reg}}[f] = \frac{1}{m}(-F^\top \mathbf{y} + F^\top F\alpha) + \lambda\alpha = 0$ and therefore $\alpha = (F^\top F + \lambda m \mathbf{1})^{-1} F^\top \mathbf{y}$

This is the same as when we added ε to the main diagonal to invert matrices or improve their condition!

A Practical Example



- Training Set
- Regression for $\lambda = 0.1$
- Regression for $\lambda = 1$
- Regression for $\lambda = 10$